More Dynamic Programming

Lecture 9

September 30, 2014

What is the running time of the following?

Consider computing $\mathbf{f(x, y)}$ by recursive function + memoization.

$$\mathbf{f(x, y)} = \sum_{i=1}^{x+y-1} x * f(x + y - i, i - 1),$$
$$f(0, y) = y \qquad f(x, 0) = x.$$

The resulting algorithm when computing $\mathbf{f(n, n)}$ 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 can not be computed.

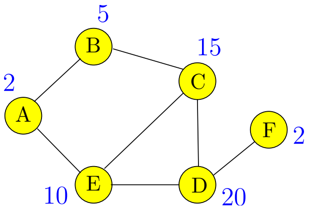
Part I

Maximum Weighted Independent Set in Trees

Maximum Weight Independent Set Problem

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

Goal Find maximum weight independent set in $\mathbf{G}$

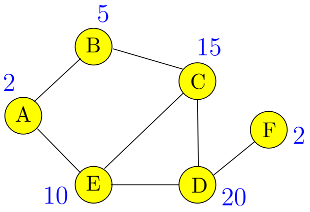


Maximum weight independent set in above graph: $\{\mathbf{B}, \mathbf{D}\}$

Maximum Weight Independent Set Problem

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

Goal Find maximum weight independent set in $\mathbf{G}$

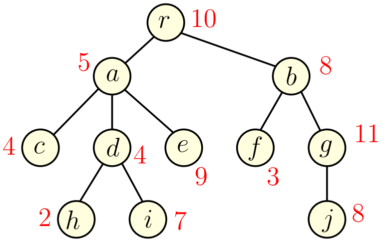


Maximum weight independent set in above graph: $\{\mathbf{B}, \mathbf{D}\}$

Maximum Weight Independent Set in a Tree

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

Goal Find maximum weight independent set in $\mathbf{T}$



Maximum weight independent set in above tree: ??

Independent set in a tree...

In a tree with $\mathbf{n}$ 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 $\mathbf{G}$:

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

What about a tree? Natural candidate for $\mathbf{v_n}$ is root $\mathbf{r}$ of $\mathbf{T}$?

Towards a Recursive Solution

For an arbitrary graph $\mathbf{G}$:

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

What about a tree? Natural candidate for $\mathbf{v_n}$ is root $\mathbf{r}$ of $\mathbf{T}$?

Towards a Recursive Solution

For an arbitrary graph $\mathbf{G}$:

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

What about a tree? Natural candidate for $\mathbf{v_n}$ is root $\mathbf{r}$ of $\mathbf{T}$?

Towards a Recursive Solution

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

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

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

Subproblems? Subtrees of $\mathbf{T}$ hanging at nodes in $\mathbf{T}$.

Towards a Recursive Solution

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

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

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

Subproblems? Subtrees of $\mathbf{T}$ hanging at nodes in $\mathbf{T}$.

Towards a Recursive Solution

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

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

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

Subproblems? Subtrees of $\mathbf{T}$ hanging at nodes in $\mathbf{T}$.

A Recursive Solution

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

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

A Recursive Solution

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

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

Iterative Algorithm

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

Iterative Algorithm

1. Compute $\mathbf{OPT(u)}$ bottom up. To evaluate $\mathbf{OPT(u)}$ need to have computed values of all children and grandchildren of $\mathbf{u}$
2. What is an ordering of nodes of a tree $\mathbf{T}$ 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**

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

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.

Iterative Algorithm

MIS-Tree(**T**):

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

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

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.

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**

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

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.

Iterative Algorithm

MIS-Tree(**T**):

```
Let v_1, v_2, ..., v_n be a post-order traversal of nodes of T
for i = 1 to n do
```

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

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.

Iterative Algorithm

MIS-Tree($\mathbf{T}$):

Let $\mathbf{v_1, v_2, \ldots, v_n}$ be a post-order traversal of nodes of T
for $\mathbf{i = 1}$ to $\mathbf{n}$ **do**

$$\mathbf{M[v_i] = \max\begin{pmatrix} \sum_{v_j \text{ child of } v_i} M[v_j], \\ w(v_i) + \sum_{v_j \text{ grandchild of } v_i} M[v_j] \end{pmatrix}}$$

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

Space: $\mathbf{O(n)}$ to store the value at each node of $\mathbf{T}$

Running time:

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

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**

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

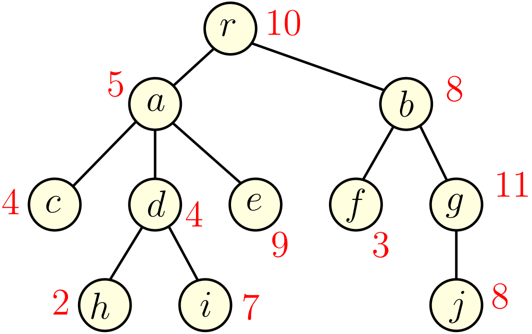
return $M[v_n]$ (* Note: v_n is the root of T *)

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.

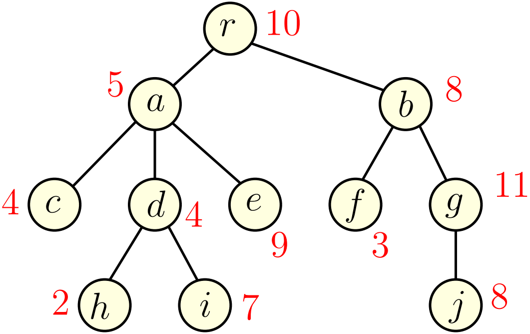
Example



Dominating set

Definition

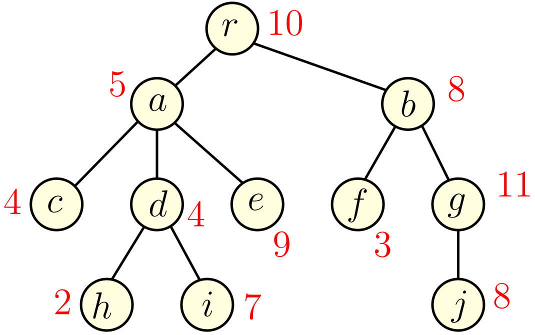
$G = (V, E)$. The set $X \subseteq V$ is a **dominating set**, if any vertex $v \in V$ is either in X or is adjacent to a vertex in X.



Dominating set

Definition

$G = (V, E)$. The set $\mathbf{X} \subseteq V$ is a **dominating set**, if any vertex $\mathbf{v} \in V$ is either in $\mathbf{X}$ or is adjacent to a vertex in $\mathbf{X}$.



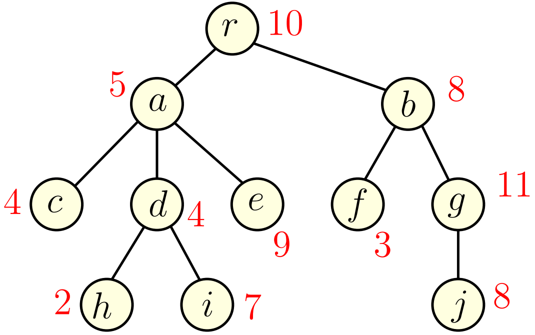
Problem

Given weights on vertices, compute the **minimum** *weight dominating set in G.*

Dominating set

Definition

$G = (V, E)$. The set $\mathbf{X} \subseteq V$ is a **dominating set**, if any vertex $\mathbf{v} \in V$ is either in $\mathbf{X}$ or is adjacent to a vertex in $\mathbf{X}$.



Problem

Given weights on vertices, compute the **minimum** *weight dominating set in G.*

Dominating Set is **NP-Hard**!

Minimum weight dominating set in a tree can be computed using the recursive formula...

(A) $\mathcal{O}(\mathbf{u}) = \min \begin{cases} \sum_{\mathbf{v}\text{ child of }\mathbf{u}} \mathcal{O}(\mathbf{v}), \\ \mathbf{w}(\mathbf{u}) + \sum_{\mathbf{v}\text{ grandchild of }\mathbf{u}} \mathcal{O}(\mathbf{v}) \end{cases}.$

(B) $\mathcal{O}(\mathbf{u}) = \mathbf{w}(\mathbf{u}) + \min \begin{cases} \sum_{\mathbf{v}\text{ child of }\mathbf{u}} \mathcal{O}(\mathbf{v}), \\ \sum_{\mathbf{v}\text{ grandchild of }\mathbf{u}} \mathcal{O}(\mathbf{v}) \end{cases}.$

(C) $\mathcal{O}(\mathbf{u}) = \mathbf{w}(\mathbf{u}) + \max \begin{cases} \sum_{\mathbf{v}\text{ child of }\mathbf{u}} \mathcal{O}(\mathbf{v}), \\ \sum_{\mathbf{v}\text{ grandchild of }\mathbf{u}} \mathcal{O}(\mathbf{v}) \end{cases}.$

(D) $\mathcal{O}(\mathbf{u}) = \mathbf{w}(\mathbf{u}) + \sum_{\mathbf{v}\text{ grandchild of }\mathbf{u}} \mathcal{O}(\mathbf{v}).$

(E) None of the above.

Part II

DAGs and Dynamic Programming

Recursion and DAGs

Observation

Let $\mathbf{A}$ *be a recursive algorithm for problem* $\mathbf{\Pi}$*. For each instance* $\mathbf{I}$ *of* $\mathbf{\Pi}$ *there is an associated* DAG $\mathbf{G(I)}$*.*

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

Recursion and DAGs

Observation

Let $\mathbf{A}$ *be a recursive algorithm for problem* $\mathbf{\Pi}$*. For each instance* $\mathbf{I}$ *of* $\mathbf{\Pi}$ *there is an associated* DAG $\mathbf{G(I)}$*.*

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

Observation

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.*

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

Let $\mathbf{O_i}$ be value of an optimal schedule for the first $\mathbf{i}$ jobs.

```
Schedule(n):
        if n = 0 then return 0
        if n = 1 then return w(v_1)
        O_p(n) ← Schedule(p(n))
        O_(n−1) ← Schedule(n − 1)
        if (O_p(n) + w(v_n) < O_(n−1)) then
            O_n = O_(n−1)
        else
            O_n = O_p(n) + w(v_n)
        return O_n
```

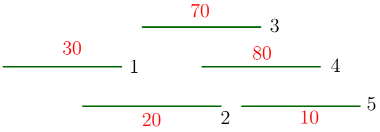
Weighted Interval Scheduling via...

Longest Path in a DAG

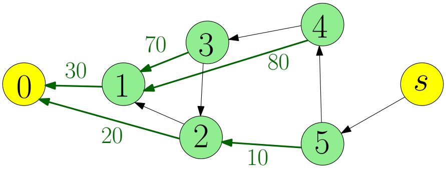
Given intervals, create a DAG as follows:

1. Create one node for each interval, plus a dummy sink node $\mathbf{0}$ for interval $\mathbf{0}$, plus a dummy source node $\mathbf{s}$.
2. For each interval $\mathbf{i}$ add edge $(\mathbf{i}, \mathbf{p}(\mathbf{i}))$ of the length/weight of $\mathbf{v_i}$.
3. Add an edge from $\mathbf{s}$ to $\mathbf{n}$ of length $\mathbf{0}$.
4. For each interval $\mathbf{i}$ add edge $(\mathbf{i}, \mathbf{i}-\mathbf{1})$ of length $\mathbf{0}$.

Example



$p(5) = 2,\ p(4) = 1,\ p(3) = 1,\ p(2) = 0,\ p(1) = 0$



Relating Optimum Solution

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

Claim

Optimum solution to weighted interval scheduling instance $\mathbf{I}$ *is given by* longest *path from* $\mathbf{s}$ *to* $\mathbf{0}$ *in* $\mathbf{G(I)}$.

Assuming claim is true,

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

Relating Optimum Solution

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

Claim

Optimum solution to weighted interval scheduling instance $\mathbf{I}$ *is given by* longest *path from* $\mathbf{s}$ *to* $\mathbf{0}$ *in* $\mathbf{G(I)}$.

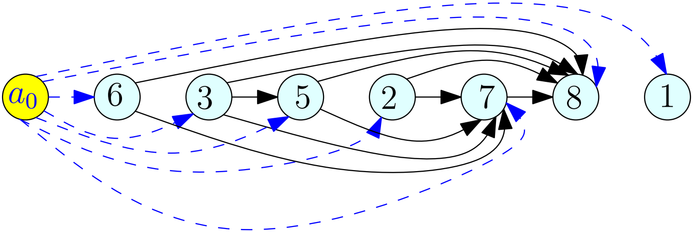
Assuming claim is true,

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

DAG for Longest Increasing Sequence

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

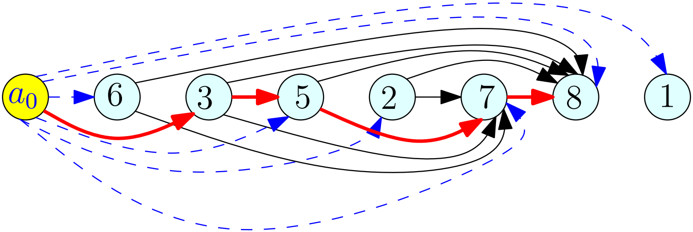
1. add sentinel a_0 to sequence where a_0 is less than smallest element in sequence
2. for each i there is a node v_i
3. if $i < j$ and $a_i < a_j$ add an edge (v_i, v_j)
4. find longest path from v_0



DAG for Longest Increasing Sequence

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

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



Part III

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?

What does nearness mean?

Question: Given two strings $\mathbf{x_1 x_2 \ldots x_n}$ and $\mathbf{y_1 y_2 \ldots y_m}$ what is a *distance* between them?

Edit Distance: minimum number of "edits" to transform $\mathbf{x}$ into $\mathbf{y}$.

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 $\mathbf{x_1 x_2 \ldots x_n}$ and $\mathbf{y_1 y_2 \ldots y_m}$ what is a *distance* between them?

Edit Distance: minimum number of "edits" to transform $\mathbf{x}$ into $\mathbf{y}$.

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 $\mathbf{x_1 x_2 \ldots x_n}$ and $\mathbf{y_1 y_2 \ldots y_m}$ what is a *distance* between them?

Edit Distance: minimum number of "edits" to transform $\mathbf{x}$ into $\mathbf{y}$.

Edit Distance

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**:

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

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.

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.

$$\begin{array}{ccccc} \mathbf{F} & \mathbf{O} & \mathbf{O} & & \mathbf{D} \\ \mathbf{M} & \mathbf{O} & \mathbf{N} & \mathbf{E} & \mathbf{Y} \end{array}$$

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

Edit Distance: 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.

$$\begin{array}{ccccc} \mathbf{F} & \mathbf{O} & \mathbf{O} & & \mathbf{D} \\ \mathbf{M} & \mathbf{O} & \mathbf{N} & \mathbf{E} & \mathbf{Y} \end{array}$$

Formally, an alignment is a set $\mathbf{M}$ of pairs $(\mathbf{i}, \mathbf{j})$ such that each index appears at most once, and there is no "crossing": $\mathbf{i} < \mathbf{i}'$ and $\mathbf{i}$ is matched to $\mathbf{j}$ implies $\mathbf{i}'$ is matched to $\mathbf{j}' > \mathbf{j}$. In the above example, this is $\mathbf{M} = \{(1,1), (2,2), (3,3), (4,5)\}$. 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 $\mathbf{X}$ and $\mathbf{Y}$, the cost of alignment $\mathbf{M}$ is

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

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

Similarity Metric

Definition

For two strings $\mathbf{X}$ and $\mathbf{Y}$, the cost of alignment $\mathbf{M}$ is

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

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

An Example

Example

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 $= 19\delta$.

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?

$\boxed{\text{SURFING}}$

$\boxed{\text{STUDYING}}$

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

Sequence Alignment

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

Goal Find alignment of minimum cost

Edit distance

Basic observation

Let $\mathbf{X} = \alpha\mathbf{x}$ and $\mathbf{Y} = \beta\mathbf{y}$

α, β: strings.

$\mathbf{x}$ and $\mathbf{y}$ single characters.

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

α	$\mathbf{x}$
β	$\mathbf{y}$

or

α	$\mathbf{x}$
$\beta\mathbf{y}$	

or

$\alpha\mathbf{x}$	
β	$\mathbf{y}$

Observation

Prefixes must have optimal alignment!

Problem Structure

Observation

Let $\mathbf{X} = \mathbf{x_1 x_2 \cdots x_m}$ and $\mathbf{Y} = \mathbf{y_1 y_2 \cdots y_n}$. If $(\mathbf{m}, \mathbf{n})$ are not matched then either the $\mathbf{m}$th position of $\mathbf{X}$ remains unmatched or the $\mathbf{n}$th position of $\mathbf{Y}$ remains unmatched.

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

Subproblems and Recurrence

Optimal Costs

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

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

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

Subproblems and Recurrence

Optimal Costs

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

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

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

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

1. Running time is $O(mn)$.

Dynamic Programming Solution

```
for all i do M[i,0] = iδ
for all j do M[0,j] = jδ

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

1. 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

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

Analysis

1. Running time is $O(mn)$.
2. Space used is $O(mn)$.

Matrix and DAG of Computation

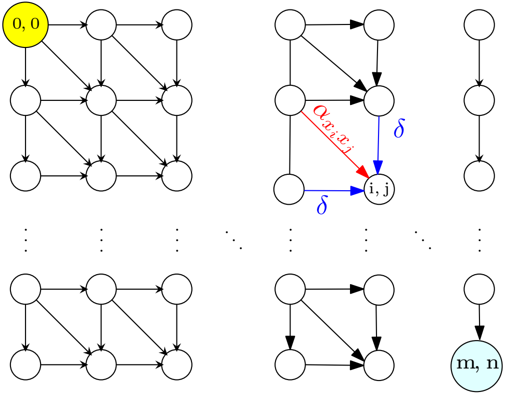


Figure : Iterative algorithm in previous slide computes values in row order. Optimal value is a shortest path from $\mathbf{(0,0)}$ to $\mathbf{(m,n)}$ in .

Sequence Alignment in Practice

1. Typically the DNA sequences that are aligned are about $\mathbf{10^5}$ letters long!
2. So about $\mathbf{10^{10}}$ operations and $\mathbf{10^{10}}$ bytes needed
3. The killer is the 10GB storage
4. Can we reduce space requirements?

Optimizing Space

1. Recall

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

2. Entries in $\mathbf{j}$th column only depend on $\mathbf{(j-1)}$st column and earlier entries in $\mathbf{j}$th column
3. Only store the current column and the previous column reusing space; $\mathbf{N(i,0)}$ stores $\mathbf{M(i,j-1)}$ and $\mathbf{N(i,1)}$ stores $\mathbf{M(i,j)}$

Computing in column order to save space

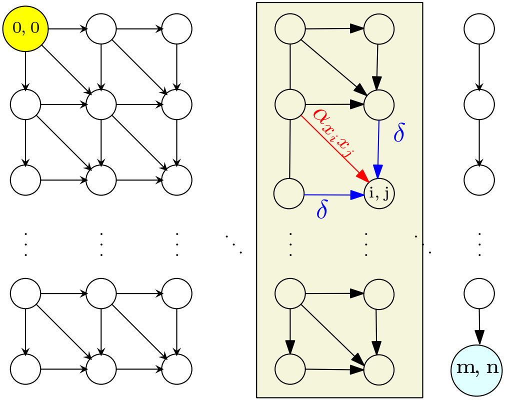


Figure : **M(i, j)** only depends on previous column values. Keep only two columns and compute in column order.

Space Efficient Algorithm

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

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

    for $\mathbf{i = 1}$ to $\mathbf{m}$ **do**
        Copy $\mathbf{N[i,0] = N[i,1]}$

Analysis

Running time is $\mathbf{O(mn)}$ and space used is $\mathbf{O(2m) = O(m)}$

Analyzing Space Efficiency

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

Takeaway Points

1. Dynamic programming is based on finding a recursive way to solve the problem. Need a recursion that generates a small number of subproblems.
2. Given a recursive algorithm there is a natural DAG associated with the subproblems that are generated for given instance; this is the dependency graph. An iterative algorithm simply evaluates the subproblems in some topological sort of this DAG.
3. The space required to evaluate the answer can be reduced in some cases by a careful examination of that dependency DAG of the subproblems and keeping only a subset of the DAG at any time.

Notes

Notes

Notes

Notes