More Dynamic Programming

Lecture 14

Tuesday, October 11, 2022

LaTeXed: October 18, 2022 13:33

14.1
Review of dynamic programming and some new problems

What is the running time of the following?

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

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

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 can not 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
   2.1 Estimate number of different subproblems generated as a function of n. Is it polynomial or exponential in n?
   2.2 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.

14.1.1
Is in L^k?

A variation

Input: A string $w \in \Sigma^*$ and access to a language $L \subseteq \Sigma^*$ via function **IsStringinL**$(string\ \ x)$ that decides whether x is in L, and non-negative integer k

Goal: Decide if $w \in L^k$ using **IsStringinL**$(string\ \ x)$ as a black box sub-routine

Example 14.1.

Suppose L is *English* and we have a procedure to check whether a string/word is in the *English* dictionary.

- Is the string "isthisanenglishsentence" in $English^5$?
- Is the string "isthisanenglishsentence" in $English^4$?
- Is "asinineat" in $English^2$?
- Is "asinineat" in $English^4$?
- Is "zibzzzad" in $English^1$?

Recursive Solution

When is $w \in L^k$?

$k = 0$: $w \in L^k$ iff $w = \epsilon$

$k = 1$: $w \in L^k$ iff $w \in L$

$k > 1$: $w \in L^k$ if $w = uv$ with $u \in L^{k-1}$ and $v \in L$

Assume w is stored in array $A[1..n]$

```
IsStringinLk(A[1 ... i], k):
    if k = 0 and i = 0 then return YES
    if k = 0 then return NO   // i > 0
    if k = 1 then
        return IsStringinL(A[1 ... i])

    for ℓ = 1 ... i − 1 do
        if IsStringinLk(A[1 ... ℓ], k − 1) and IsStringinL(A[ℓ + 1 ... i]) then
            return YES

    return NO
```

Recursive Solution

When is $w \in L^k$?

$k = 0$: $w \in L^k$ iff $w = \epsilon$

$k = 1$: $w \in L^k$ iff $w \in L$

$k > 1$: $w \in L^k$ if $w = uv$ with $u \in L^{k-1}$ and $v \in L$

Assume w is stored in array $A[1..n]$

```
IsStringinLk(A[1...i], k):
    if k = 0 and i = 0 then return YES
    if k = 0 then return NO   // i > 0
    if k = 1 then
        return IsStringinL(A[1...i])

    for ℓ = 1...i − 1 do
        if IsStringinLk(A[1...ℓ], k − 1) and IsStringinL(A[ℓ + 1...i]) then
            return YES

    return NO
```

Recursive Solution

When is $w \in L^k$?

$k = 0$: $w \in L^k$ iff $w = \epsilon$

$k = 1$: $w \in L^k$ iff $w \in L$

$k > 1$: $w \in L^k$ if $w = uv$ with $u \in L^{k-1}$ and $v \in L$

Assume w is stored in array $A[1..n]$

```
IsStringinLk(A[1 ... i], k):
    if k = 0 and i = 0 then return YES
    if k = 0 then return NO   // i > 0
    if k = 1 then
        return IsStringinL(A[1 ... i])

    for ℓ = 1 ... i − 1 do
        if IsStringinLk(A[1 ... ℓ], k − 1) and IsStringinL(A[ℓ + 1 ... i]) then
            return YES

    return NO
```

Analysis

```
IsStringinLk(A[1...i], k):
    if k = 0 and i = 0 then return YES
    if k = 0 then return NO    // i > 0
    if k = 1 then
        return IsStringinL(A[1...i])

    for ℓ = 1...i − 1 do
        if IsStringinLk(A[1...ℓ], k − 1) and IsStringinL(A[ℓ + 1...i]) then
            return YES

    return NO
```

- How many distinct sub-problems are generated by **IsStringinLk**$(A[1..n], k)$? $O(nk)$
- How much space? $O(nk)$
- Running time if we use memoization? $O(n^2k)$

Analysis

```
IsStringinLk(A[1...i], k):
    if k = 0 and i = 0 then return YES
    if k = 0 then return NO   // i > 0
    if k = 1 then
        return IsStringinL(A[1...i])

    for ℓ = 1...i − 1 do
        if IsStringinLk(A[1...ℓ], k − 1) and IsStringinL(A[ℓ + 1...i]) then
            return YES

    return NO
```

- How many distinct sub-problems are generated by **IsStringinLk**$(A[1..n], k)$? $O(nk)$
- How much space? $O(nk)$
- Running time if we use memoization? $O(n^2 k)$

Analysis

```
IsStringinLk(A[1...i], k):
    if k = 0 and i = 0 then return YES
    if k = 0 then return NO   // i > 0
    if k = 1 then
        return IsStringinL(A[1...i])

    for ℓ = 1...i − 1 do
        if IsStringinLk(A[1...ℓ], k − 1) and IsStringinL(A[ℓ + 1...i]) then
            return YES

    return NO
```

- How many distinct sub-problems are generated by **IsStringinLk**$(A[1..n], k)$? $O(nk)$
- How much space? $O(nk)$
- Running time if we use memoization? $O(n^2 k)$

Analysis

```
IsStringinLk(A[1...i], k):
    if k = 0 and i = 0 then return YES
    if k = 0 then return NO   // i > 0
    if k = 1 then
        return IsStringinL(A[1...i])

    for ℓ = 1...i − 1 do
        if IsStringinLk(A[1...ℓ], k − 1) and IsStringinL(A[ℓ + 1...i]) then
            return YES

    return NO
```

- How many distinct sub-problems are generated by **IsStringinLk**$(A[1..n], k)$? $O(nk)$
- How much space? $O(nk)$
- Running time if we use memoization? $O(n^2k)$

Analysis

```
IsStringinLk(A[1...i], k):
    if k = 0 and i = 0 then return YES
    if k = 0 then return NO   // i > 0
    if k = 1 then
        return IsStringinL(A[1...i])

    for ℓ = 1...i − 1 do
        if IsStringinLk(A[1...ℓ], k − 1) and IsStringinL(A[ℓ + 1...i]) then
            return YES

    return NO
```

- How many distinct sub-problems are generated by **IsStringinLk**$(A[1..n], k)$? $O(nk)$
- How much space? $O(nk)$
- Running time if we use memoization? $O(n^2 k)$

Analysis

```
IsStringinLk(A[1 ... i], k):
    if k = 0 and i = 0 then return YES
    if k = 0 then return NO   // i > 0
    if k = 1 then
        return IsStringinL(A[1 ... i])

    for ℓ = 1 ... i − 1 do
        if IsStringinLk(A[1 ... ℓ], k − 1) and IsStringinL(A[ℓ + 1 ... i]) then
            return YES

    return NO
```

- How many distinct sub-problems are generated by **IsStringinLk**$(A[1..n], k)$? $O(nk)$
- How much space? $O(nk)$
- Running time if we use memoization? $O(n^2 k)$

Another variant

Question: What if we want to check if $w \in L^i$ for some $0 \leq i \leq k$? That is, is $w \in \cup_{i=0}^{k} L^i$?

Exercise

Definition 14.2.

A string is a palindrome if $w = w^R$.
Examples: I, $RACECAR$, $MALAYALAM$, $DOOFFOOD$

Problem: Given a string w find the longest subsequence of w that is a palindrome.

Example 14.3.

$MAHDYNAMICPROGRAMZLETMESHOWYOUTHEM$ has $MHYMRORMYHM$ as a palindromic subsequence

Exercise

Definition 14.2.

A string is a palindrome if $w = w^R$.
Examples: I, $RACECAR$, $MALAYALAM$, $DOOFFOOD$

Problem: Given a string w find the longest subsequence of w that is a palindrome.

Example 14.3.

$MAHDYNAMICPROGRAMZLETMESHOWYOUTHEM$ has $MHYMRORMYHM$ as a palindromic subsequence

Exercise

Assume w is stored in an array $A[1..n]$

$LPS(A[1..n])$: length of longest palindromic subsequence of A.

Recursive expression/code?

14.2
Edit Distance and Sequence Alignment

14.2.1
Problem definition and background

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

Edit Distance: minimum number of "edits" to transform x into 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 $x_1 x_2 \ldots x_n$ and $y_1 y_2 \ldots y_m$ what is a distance between them?

Edit Distance: minimum number of "edits" to transform x into 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 $x_1 x_2 \ldots x_n$ and $y_1 y_2 \ldots y_m$ what is a distance between them?

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

Edit Distance

Definition 14.1.

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 14.2.

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 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 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 14.3.

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 δ.
2. [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 $\delta = \alpha_{pq} = 1$.

Similarity Metric

Definition 14.3.

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 δ.
2. [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 $\delta = \alpha_{pq} = 1$.

14.2.2
Edit distance as alignment

An Example

Example 14.4.

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|} o & & c & u & r & r & a & n & c & e \\ o & c & c & u & r & r & e & n & c & e \end{array} \qquad \text{Cost } = \delta + \alpha_{ae}$$

Alternative:

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c} o & & c & u & r & r & & a & n & c & e \\ o & c & c & u & r & r & e & & n & c & e \end{array} \qquad \text{Cost } = 3\delta$$

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

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c} o & c & u & r & r & a & n & c & e & & & & & & & & & & \\ & & & & & & & & & o & c & c & u & r & r & e & n & c & e \end{array}$$

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?

$\boxed{374}$

$\boxed{473}$

(A) 1

(B) 2

(C) 3

(D) 4

(E) 5

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?

$\boxed{373}$

$\boxed{473}$

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

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?

$\boxed{37}$

$\boxed{473}$

(A) 1

(B) 2

(C) 3

(D) 4

(E) 5

Sequence Alignment

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

Goal Find alignment of minimum cost

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?

14.2.3
Edit distance: The algorithm

Edit distance

Basic observation

Let $X = \alpha x$ and $Y = \beta y$

α, β: 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:

α	x
β	y

or

α	x
βy	

or

αx	
β	y

Observation 14.5.

Prefixes must have optimal alignment!

Problem Structure

Observation 14.6.

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

1. Case x_m and y_n are matched.
   1.1 Pay mismatch cost $\alpha_{x_m y_n}$ plus cost of aligning strings $x_1 \cdots x_{m-1}$ and $y_1 \cdots y_{n-1}$
2. Case x_m is unmatched.
   2.1 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.
   3.1 Pay gap penalty plus cost of aligning $x_1 \cdots x_m$ and $y_1 \cdots y_{n-1}$

Subproblems and Recurrence

$x_1 \ldots x_{i-1}$	x_i	or	$x_1 \ldots x_{i-1}$	x	or	$x_1 \ldots x_{i-1} x_i$	
$y_1 \ldots y_{j-1}$	y_j		$y_1 \ldots y_{j-1} y_j$			$y_1 \ldots y_{j-1}$	y_j

Optimal Costs

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

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

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

Subproblems and Recurrence

$x_1 \ldots x_{i-1}$	x_i	or	$x_1 \ldots x_{i-1}$	x	or	$x_1 \ldots x_{i-1} x_i$	
$y_1 \ldots y_{j-1}$	y_j		$y_1 \ldots y_{j-1} y_j$			$y_1 \ldots y_{j-1}$	y_j

Optimal Costs

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

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

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

Recursive Algorithm

Assume X is stored in array $A[1..m]$ and Y is stored in $B[1..n]$
Array $COST$ stores cost of matching two chars. Thus $COST[a, b]$ give the cost of matching character a to character b.

```
EDIST(A[1..m], B[1..n])
    If (m = 0) return nδ
    If (n = 0) return mδ
    m1 = δ + EDIST(A[1..(m − 1)], B[1..n])
    m2 = δ + EDIST(A[1..m], B[1..(n − 1)]))
    m3 = COST[A[m], B[n]] + EDIST(A[1..(m − 1)], B[1..(n − 1)])
    return min(m1, m2, m3)
```

Example: DEED and DREAD

	ε	D	R	E	A	D
ε						
D						
E						
E						
D						

Example: DEED and DREAD

	ε	*D*	*R*	*E*	*A*	*D*
ε	0	1	2	3	4	5
D	1					
E	2					
E	3					
D	3					

Example: DEED and DREAD

	ε	D	R	E	A	D
ε	0	1	2	3	4	5
D	1	0	1	2	3	4
E	2					
E	3					
D	3					

Example: DEED and DREAD

	ε	*D*	*R*	*E*	*A*	*D*
ε	0	1	2	3	4	5
D	1	0	1	2	3	4
E	2	1	1	1	2	3
E	3					
D	3					

Example: DEED and DREAD

	ε	D	R	E	A	D
ε	0	1	2	3	4	5
D	1	0	1	2	3	4
E	2	1	1	1	2	3
E	3	2	2	1	2	3
D	3					

Example: DEED and DREAD

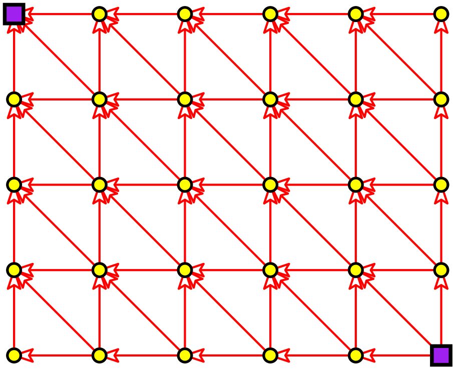
	ε	*D*	*R*	*E*	*A*	*D*
ε	0	1	2	3	4	5
D	1	0	1	2	3	4
E	2	1	1	1	2	3
E	3	2	2	1	2	3
D	3	3	3	2	2	2

D	R	E	A	D
D	E	E		D

Example: DEED and DREAD

	ε	*D*	*R*	*E*	*A*	*D*
ε	0	1	2	3	4	5
D	1	0	1	2	3	4
E	2	1	1	1	2	3
E	3	2	2	1	2	3
D	3	3	3	2	2	2

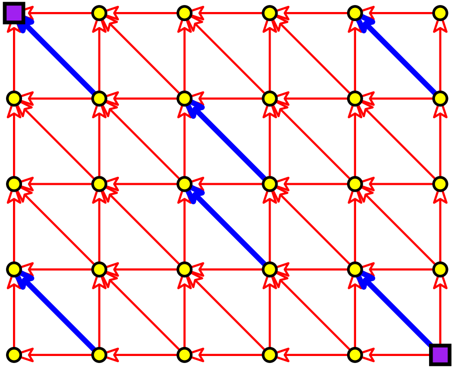
D	R	E	A	D
D	E	E		D



Example: DEED and DREAD

	ε	*D*	*R*	*E*	*A*	*D*
ε	0	1	2	3	4	5
D	1	0	1	2	3	4
E	2	1	1	1	2	3
E	3	2	2	1	2	3
D	3	3	3	2	2	2

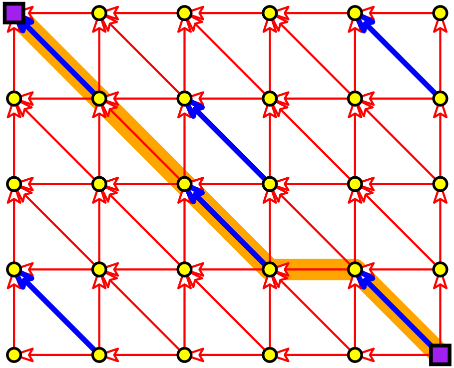
D	R	E	A	D
D	E	E		D



Example: DEED and DREAD

	ε	*D*	*R*	*E*	*A*	*D*
ε	0	1	2	3	4	5
D	1	0	1	2	3	4
E	2	1	1	1	2	3
E	3	2	2	1	2	3
D	3	3	3	2	2	2

D	R	E	A	D
D	E	E		D



14.2.4
Dynamic programming algorithm for edit-distance

As part of the input...

The cost of aligning a character against another character

$\boldsymbol{\Sigma}$: Alphabet

We are given a **cost** function (in a table):

$$\forall b, c \in \boldsymbol{\Sigma} \qquad COST[b][c] = \text{ cost of aligning } b \text{ with } c.$$
$$\forall b \in \boldsymbol{\Sigma} \qquad COST[b][b] = 0$$

δ : price of deletion of insertion of a single character

Memoizing the Recursive Algorithm (Explicit Memoization)

Input: Two strings
$A[1 \ldots m]$
$B[1 \ldots n]$

```
EditDistance(A, B)
    int  M[0..m][0..n]
    ∀i, j  M[i][j] ← ∞
    return edEMI(m, n)
```

```
edEMI( i , j )        // A[1 ... i], B[1 ... j]
    if M[i][j] < ∞
        return M[i][j]    // stored value

    if i = 0 or j = 0
        M[i][j] = (i + j)δ
        return M[i][j]

    m1 = δ + edEMI( i − 1 , j )
    m2 = δ + edEMI( i , j − 1)

    m3 = COST[A[i]][B[j]]
         +  edEMI( i − 1 , j − 1 )

    M[i][j] = min(m1, m2, m3)
    return M[i][j]
```

Dynamic program for edit distance

Removing Recursion to obtain Iterative Algorithm

$EDIST(A[1..m], B[1..n])$

$\quad int \quad M[0..m][0..n]$

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

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

$\quad$ **for** $i = 1$ `to` m **do**

$\quad\quad$ **for** $j = 1$ `to` n **do**

$$M[i][j] = \min \begin{cases} COST\left[A[i]\right]\left[B[j]\right] + M[i-1][j-1], \\ \delta + M[i-1][j], \\ \delta + M[i][j-1] \end{cases}$$

Analysis

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

Dynamic program for edit distance

Removing Recursion to obtain Iterative Algorithm

```
EDIST(A[1..m], B[1..n])
    int  M[0..m][0..n]
    for i = 1 to m do M[i,0] = iδ
    for j = 1 to n do M[0,j] = jδ

    for i = 1 to m do
        for j = 1 to n do
```

$$M[i][j] = \min \begin{cases} COST\left[A[i]\right]\left[B[j]\right] + M[i-1][j-1], \\ \delta + M[i-1][j], \\ \delta + M[i][j-1] \end{cases}$$

Analysis

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

Dynamic program for edit distance

Removing Recursion to obtain Iterative Algorithm

```
EDIST(A[1..m], B[1..n])
    int  M[0..m][0..n]
    for i = 1 to m do M[i,0] = iδ
    for j = 1 to n do M[0,j] = jδ

    for i = 1 to m do
        for j = 1 to n do
```

$$M[i][j] = \min \begin{cases} COST\left[A[i]\right]\left[B[j]\right] + M[i-1][j-1], \\ \delta + M[i-1][j], \\ \delta + M[i][j-1] \end{cases}$$

Analysis

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

14.2.5
Reducing space for edit distance

Matrix and DAG of computation of edit distance

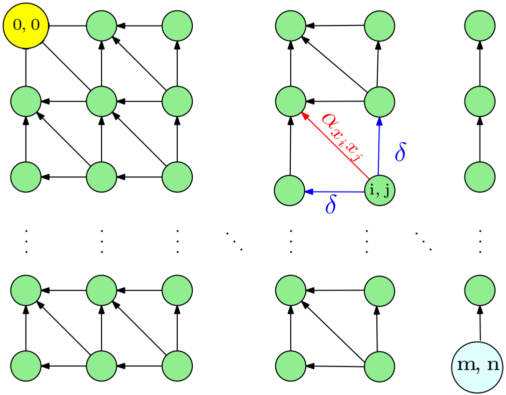


Figure: Iterative algorithm in previous slide computes values in row order.

Optimizing Space

1. Recall
$$M(i,j) = \min \begin{cases} \alpha_{x_i y_j} + M(i-1, j-1), \\ \delta + M(i-1, j), \\ \delta + M(i, j-1) \end{cases}$$
2. Entries in jth column only depend on $(j-1)$st column and earlier entries in jth column
3. Only store the current column and the previous column reusing space; $N(i, 0)$ stores $M(i, j-1)$ and $N(i, 1)$ stores $M(i, j)$

Example: DEED vs. DREAD filled by column

	ε	D	R	E	A	D
ε						
D						
E						
E						
D						

Example: DEED vs. DREAD filled by column

	ε	*D*	*R*	*E*	*A*	*D*
ε	0	1	2	3	4	5
D	1					
E	2					
E	3					
D	4					

Example: DEED vs. DREAD filled by column

	ε	**D**	**R**	**E**	**A**	**D**
ε	0	1	2	3	4	5
D	1	0				
E	2	1				
E	3	2				
D	4	3				

Example: DEED vs. DREAD filled by column

	ε	*D*	*R*	*E*	*A*	*D*
ε	0	1	2	3	4	5
D	1	0	1			
E	2	1	1			
E	3	2	2			
D	4	3	3			

Example: DEED vs. DREAD filled by column

	ε	*D*	*R*	*E*	*A*	*D*
ε	0	1	2	3	4	5
D	1	0	1	2		
E	2	1	1	1		
E	3	2	2	1		
D	4	3	3	2		

Example: DEED vs. DREAD filled by column

	ε	D	R	E	A	D
ε	0	1	2	3	4	5
D	1	0	1	2	3	
E	2	1	1	1	2	
E	3	2	2	1	2	
D	4	3	3	2	2	

Example: DEED vs. DREAD filled by column

	ε	D	R	E	A	D
ε	0	1	2	3	4	5
D	1	0	1	2	3	4
E	2	1	1	1	2	3
E	3	2	2	1	2	3
D	4	3	3	2	2	2

Computing in column order to save space

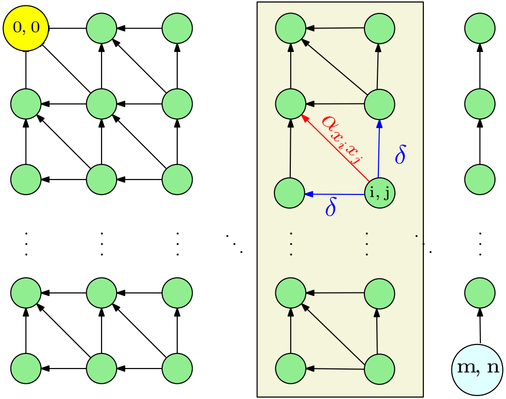


Figure: $\boldsymbol{M(i,j)}$ only depends on previous column values. Keep only two columns and compute in column order.

Space Efficient Algorithm

```
for all i do N[i,0] = iδ
for j = 1 to n do
    N[0,1] = jδ (* 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
        Copy N[i,0] = N[i,1]
```

Analysis

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

Analyzing Space Efficiency

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
3. Space efficient computation of alignment? More complicated algorithm — see notes and Kleinberg-Tardos book.

14.2.6
Longest Common Subsequence Problem

LCS Problem

Definition 14.7.
LCS between two strings X and Y is the length of longest common subsequence between X and Y.

$$\begin{array}{l} ABAZDC \\ BACBAD \end{array} \qquad\qquad \begin{array}{l} ABAZDC \\ BACBAD \end{array}$$

Example 14.8.
LCS between ABAZDC and BACBAD is 4 via ABAD

Derive a dynamic programming algorithm for the problem.

LCS Problem

Definition 14.7.

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

$$\begin{array}{l} ABAZDC \\ BACBAD \end{array} \qquad\qquad \begin{array}{l} ABAZDC \\ BACBAD \end{array}$$

Example 14.8.

LCS between ABAZDC and BACBAD is 4 via ABAD

Derive a dynamic programming algorithm for the problem.

LCS Problem

Definition 14.7.

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

$ABAZDC$	$ABAZDC$
$BACBAD$	$BACBAD$

Example 14.8.

LCS between ABAZDC and BACBAD is 4 via ABAD

Derive a dynamic programming algorithm for the problem.

LCS recursive definition

$A[1..n], B[1..m]$: Input strings.

$$LCS(i,j) = \begin{cases} 0 & i = 0 \text{ or } j = 0 \\ \max\begin{pmatrix} LCS(i-1,j), \\ LCS(i,j-1) \end{pmatrix} & A[i] \neq B[j] \\ \max\begin{pmatrix} LCS(i-1,j), \\ LCS(i,j-1), \\ 1 + LCS(i-1,j-1) \end{pmatrix} & A[i] = B[j] \end{cases}$$

Similar to edit distance... $O(nm)$ time algorithm $O(m)$ space. Better recurrence with a bit of thinking:

$$LCS(i,j) = \begin{cases} 0 & i = 0 \text{ or } j = 0 \\ \max\begin{pmatrix} LCS(i-1,j), \\ LCS(i,j-1) \end{pmatrix} & A[i] \neq B[j] \\ 1 + LCS(i-1,j-1) & A[i] = B[j]. \end{cases}$$

LCS recursive definition

$A[1..n], B[1..m]$: Input strings.

$$LCS(i,j) = \begin{cases} 0 & i = 0 \text{ or } j = 0 \\ \max\begin{pmatrix} LCS(i-1,j), \\ LCS(i,j-1) \end{pmatrix} & A[i] \neq B[j] \\ \max\begin{pmatrix} LCS(i-1,j), \\ LCS(i,j-1), \\ 1 + LCS(i-1,j-1) \end{pmatrix} & A[i] = B[j] \end{cases}$$

Similar to edit distance... $O(nm)$ time algorithm $O(m)$ space. Better recurrence with a bit of thinking:

$$LCS(i,j) = \begin{cases} 0 & i = 0 \text{ or } j = 0 \\ \max\begin{pmatrix} LCS(i-1,j), \\ LCS(i,j-1) \end{pmatrix} & A[i] \neq B[j] \\ 1 + LCS(i-1,j-1) & A[i] = B[j]. \end{cases}$$

Longest common subsequence is just edit distance for the two sequences...

A, B: input sequences
Σ: "alphabet" all the different values in A and B

$$\forall b, c \in \Sigma : b \neq c \qquad COST[b][c] = +\infty.$$
$$\forall b \in \Sigma \qquad COST[b][b] = 1$$

1 : price of deletion of insertion of a single character

Length of longest common subsequence $= m + n - \mathrm{ed}(A, B)$

Longest common subsequence is just edit distance for the two sequences...

$\boldsymbol{A}, \boldsymbol{B}$: input sequences
$\boldsymbol{\Sigma}$: "alphabet" all the different values in $\boldsymbol{A}$ and $\boldsymbol{B}$

$$\begin{array}{ll} \forall b, c \in \Sigma : b \neq c & COST[b][c] = +\infty. \\ \forall b \in \Sigma & COST[b][b] = 1 \end{array}$$

$\mathbf{1}$: price of deletion of insertion of a single character

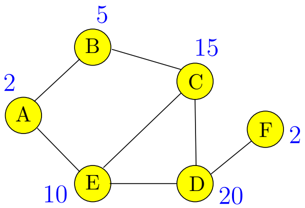
Length of longest common subsequence $= m + n - \mathrm{ed}(A, B)$

14.3
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

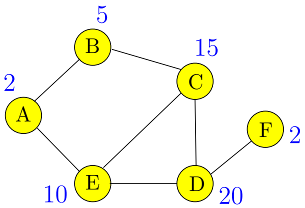


Maximum weight independent set in above graph: $\{B, D\}$

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

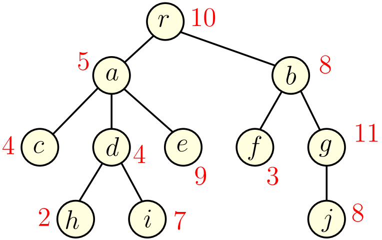


Maximum weight independent set in above graph: $\{B, D\}$

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: ??

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

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

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 rooted 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 $\mathcal{O}$ 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 rooted 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 $\mathcal{O}$ 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 rooted 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 $\mathcal{O}$ 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 rooted 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 $\mathcal{O}$ be an optimum solution to the whole problem.

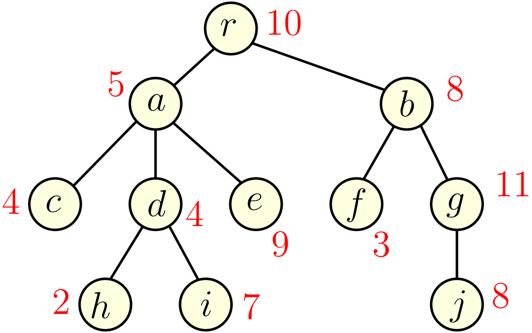
Case $r \notin \mathcal{O}$: Then $\mathcal{O}$ contains an optimum solution for each subtree of T hanging at a child of r.

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 rooted at nodes in T.

How many of them? $O(n)$

Example



A Recursive Solution

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

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

A Recursive Solution

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

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

Iterative Algorithm

1. Compute $OPT(u)$ bottom up. To evaluate $OPT(u)$ 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

1. Compute $OPT(u)$ bottom up. To evaluate $OPT(u)$ 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**

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

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

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

    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 $v_1, v_2, \ldots, v_n$ 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 *)
```

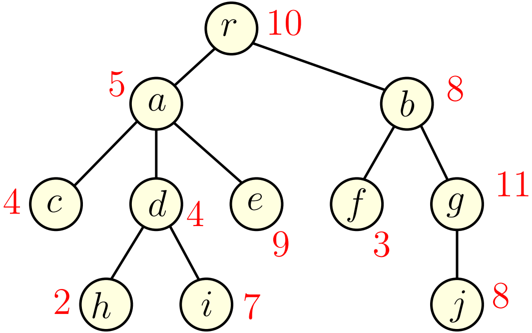
$$M[v_i] = \max\left(\begin{array}{c} \sum_{v_j \text{ child of } v_i} M[v_j], \\ w(v_i) + \sum_{v_j \text{ grandchild of } v_i} M[v_j] \end{array}\right)$$

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.

Example



14.4
Dynamic programming and DAGs

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.

14.5
Supplemental: Context free grammars: The CYK Algorithm

14.5.1
CYK: Problem statement, basic idea, and an example

Parsing

We saw regular languages and context free languages.

Most programming languages are specified via context-free grammars. Why?

- CFLs are sufficiently expressive to support what is needed.
- At the same time one can "efficiently" solve the parsing problem: given a string/program w, is it a valid program according to the CFG specification of the programming language?

CFG specification for C

```
<relational-expression> ::= <shift-expression>
                          | <relational-expression> < <shift-expression>
                          | <relational-expression> > <shift-expression>
                          | <relational-expression> <= <shift-expression>
                          | <relational-expression> >= <shift-expression>

<shift-expression> ::= <additive-expression>
                     | <shift-expression> << <additive-expression>
                     | <shift-expression> >> <additive-expression>

<additive-expression> ::= <multiplicative-expression>
                        | <additive-expression> + <multiplicative-expression>
                        | <additive-expression> - <multiplicative-expression>

<multiplicative-expression> ::= <cast-expression>
                              | <multiplicative-expression> * <cast-expression>
                              | <multiplicative-expression> / <cast-expression>
                              | <multiplicative-expression> % <cast-expression>

<cast-expression> ::= <unary-expression>
                    | ( <type-name> ) <cast-expression>

<unary-expression> ::= <postfix-expression>
                     | ++ <unary-expression>
                     | -- <unary-expression>
                     | <unary-operator> <cast-expression>
                     | sizeof <unary-expression>
                     | sizeof <type-name>

<postfix-expression> ::= <primary-expression>
                       | <postfix-expression> [ <expression> ]
                       | <postfix-expression> ( {<assignment-expression>}* )
```

Algorithmic Problem

Given a CFG $G = (V, T, P, S)$ and a string $w \in T^*$, is $w \in L(G)$?

- That is, does S derive w?
- Equivalently, is there a parse tree for w?

Simplifying assumption: G is in Chomsky Normal Form (CNF)

- Productions are all of the form $A \to BC$ or $A \to a$.
  If $\epsilon \in L$ then $S \to \epsilon$ is also allowed.
  (This is the only place in the grammar that has an ε.)
- Every CFG G can be converted into CNF form via an efficient algorithm
- Advantage: parse tree of constant degree.

Algorithmic Problem

Given a CFG $G = (V, T, P, S)$ and a string $w \in T^*$, is $w \in L(G)$?

- That is, does S derive w?
- Equivalently, is there a parse tree for w?

Simplifying assumption: G is in Chomsky Normal Form (CNF)

- Productions are all of the form $A \to BC$ or $A \to a$.
  If $\epsilon \in L$ then $S \to \epsilon$ is also allowed.
  (This is the only place in the grammar that has an ε.)
- Every CFG G can be converted into CNF form via an efficient algorithm
- Advantage: parse tree of constant degree.

Towards Recursive Algorithm

CYK Algorithm = Cocke-Younger-Kasami algorithm

Assume G is a CNF grammar.

S derives $w \iff$ one of the following holds:

- $|w| = 1$ and $S \to w$ is a rule in P
- $|w| > 1$ and there is a rule $S \to AB$ and a split $w = uv$ with $|u|, |v| \geq 1$ such that A derives u and B derives v

Observation: Subproblems generated require us to know if some non-terminal A will derive a substring of w.

Towards Recursive Algorithm

CYK Algorithm = Cocke-Younger-Kasami algorithm

Assume G is a CNF grammar.

S derives $w \iff$ one of the following holds:

- $|w| = 1$ and $S \rightarrow w$ is a rule in P
- $|w| > 1$ and there is a rule $S \rightarrow AB$ and a split $w = uv$ with $|u|, |v| \geq 1$ such that A derives u and B derives v

Observation: Subproblems generated require us to know if some non-terminal A will derive a substring of w.

Example

$S \rightarrow \epsilon \mid AB \mid XB$
$Y \rightarrow AB \mid XB$
$X \rightarrow AY$
$A \rightarrow \mathbf{0}$
$B \rightarrow \mathbf{1}$

Question:

- Is $\mathbf{000111}$ in $L(G)$?
- Is $\mathbf{00011}$ in $L(G)$?

Order of evaluation for iterative algorithm: increasing order of substring length.

Example: $\mathbf{000111}$

$S \to \epsilon \mid AB \mid XB$
$Y \to AB \mid XB$
$X \to AY$
$A \to \mathbf{0}$
$B \to \mathbf{1}$

Input:	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$

Example: $\mathbf{000111}$

$S \rightarrow \epsilon \mid AB \mid XB$
$Y \rightarrow AB \mid XB$
$X \rightarrow AY$
$A \rightarrow \mathbf{0}$
$B \rightarrow \mathbf{1}$

Len=1	A	A	A	B	B	B
Input:	**0**	**0**	**0**	**1**	**1**	**1**

Example: $\mathbf{000111}$

$S \to \epsilon \mid AB \mid XB$
$Y \to AB \mid XB$
$X \to AY$
$A \to \mathbf{0}$
$B \to \mathbf{1}$

Len=3		X				
Len=2			Y			
Len=1	A	A	A	B	B	B
Input:	**0**	**0**	**0**	**1**	**1**	**1**

Example: $\mathbf{000111}$

$S \rightarrow \epsilon \mid AB \mid XB$
$Y \rightarrow AB \mid XB$
$X \rightarrow AY$
$A \rightarrow \mathbf{0}$
$B \rightarrow \mathbf{1}$

Len=4		Y,S				
Len=3		X				
Len=2			Y			
Len=1	A	A	A	B	B	B
Input:	**0**	**0**	**0**	**1**	**1**	**1**

Example: $\mathbf{000111}$

$S \to \epsilon \mid AB \mid XB$
$Y \to AB \mid XB$
$X \to AY$
$A \to \mathbf{0}$
$B \to \mathbf{1}$

Len=5	X					
Len=4		Y,S				
Len=3		X				
Len=2			Y			
Len=1	A	A	A	B	B	B
Input:	**0**	**0**	**0**	**1**	**1**	**1**

Example: **000111**

$S \rightarrow \epsilon \mid AB \mid XB$
$Y \rightarrow AB \mid XB$
$X \rightarrow AY$
$A \rightarrow \mathbf{0}$
$B \rightarrow \mathbf{1}$

Len=6	S					
Len=5	X					
Len=4		Y,S				
Len=3		X				
Len=2			Y			
Len=1	A	A	A	B	B	B
Input:	**0**	**0**	**0**	**1**	**1**	**1**

Example II: **00111**

$S \to \epsilon \mid AB \mid XB$
$Y \to AB \mid XB$
$X \to AY$
$A \to \mathbf{0}$
$B \to \mathbf{1}$

Input:	**0**	**0**	**1**	**1**	**1**

Example II: $\mathbf{00111}$

$S \to \epsilon \mid AB \mid XB$
$Y \to AB \mid XB$
$X \to AY$
$A \to \mathbf{0}$
$B \to \mathbf{1}$

Len=1	A	A	B	B	B
Input:	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$

Example II: **00111**

$S \rightarrow \epsilon \mid AB \mid XB$
$Y \rightarrow AB \mid XB$
$X \rightarrow AY$
$A \rightarrow \mathbf{0}$
$B \rightarrow \mathbf{1}$

Len=3	X				
Len=2		Y			
Len=1	A	A	B	B	B
Input:	**0**	**0**	**1**	**1**	**1**

Example II: $\mathbf{00111}$

$S \rightarrow \epsilon \mid AB \mid XB$
$Y \rightarrow AB \mid XB$
$X \rightarrow AY$
$A \rightarrow \mathbf{0}$
$B \rightarrow \mathbf{1}$

Len=4	Y,S				
Len=3	X				
Len=2		Y			
Len=1	A	A	B	B	B
Input:	**0**	**0**	**1**	**1**	**1**

Example II: **00111**

$S \to \epsilon \mid AB \mid XB$
$Y \to AB \mid XB$
$X \to AY$
$A \to \mathbf{0}$
$B \to \mathbf{1}$

Len=5					
Len=4	Y,S				
Len=3	X				
Len=2		Y			
Len=1	A	A	B	B	B
Input:	**0**	**0**	**1**	**1**	**1**

14.5.2
Formal description of algorithm

Recursive solution

1. Input: $w = w_1 w_2 \dots w_n$
2. Assume r non-terminals in G: $R_1, \dots, R_r$.
3. R_1: Start symbol.
4. $f(\ell, s, b)$: TRUE $\iff$ $w_s w_{s+1} \dots, w_{s+\ell-1} \in L(R_b)$.
   $=$ Substring w starting at pos ℓ of length s is deriveable by R_b.
5. **Recursive formula:** $f(1, s, a)$ is $1 \iff \left(R_a \to w_s\right) \in G$.
6. For $\ell > 1$: $f\left(\text{length}, \text{start pos}, \text{variable index}\right)$

$$f(\ell, s, a) = \bigvee_{\mu=1}^{\ell-1} \bigvee_{\left(R_a \to R_\beta R_\gamma\right) \in G} \left(f(\mu, s, \beta) \wedge f(\ell - \mu, s + \mu, \gamma) \right)$$

7. **Output:** $w \in L(G) \iff f(n, 1, 1) = 1$.

Recursive solution

1. Input: $w = w_1 w_2 \ldots w_n$
2. Assume r non-terminals in G: $R_1, \ldots, R_r$.
3. R_1: Start symbol.
4. $f(\ell, s, b)$: TRUE $\iff$ $w_s w_{s+1} \ldots, w_{s+\ell-1} \in L(R_b)$.
   = Substring w starting at pos ℓ of length s is deriveable by R_b.
5. **Recursive formula:** $f(1, s, a)$ is 1 $\iff$ $\left(R_a \rightarrow w_s\right) \in G$.
6. For $\ell > 1$: $f\left(\text{length}, \text{start pos}, \text{variable index}\right)$

$$f(\ell, s, a) = \bigvee_{\mu=1}^{\ell-1} \bigvee_{\left(R_a \rightarrow R_\beta R_\gamma\right) \in G} \left(f(\mu, s, \beta) \wedge f(\ell - \mu, s + \mu, \gamma)\right)$$

7. **Output:** $w \in L(G) \iff f(n, 1, 1) = 1$.

Recursive solution

1. Input: $w = w_1 w_2 \ldots w_n$
2. Assume r non-terminals in G: $R_1, \ldots, R_r$.
3. R_1: Start symbol.
4. $f(\ell, s, b)$: TRUE $\iff$ $w_s w_{s+1} \ldots, w_{s+\ell-1} \in L(R_b)$.
   $=$ Substring w starting at pos ℓ of length s is deriveable by R_b.
5. **Recursive formula:** $f(1, s, a)$ is 1 $\iff$ $\left(R_a \to w_s\right) \in G$.
6. For $\ell > 1$: $f\left(\text{length}, \text{start pos}, \text{variable index}\right)$

$$f(\ell, s, a) = \bigvee_{\mu=1}^{\ell-1} \bigvee_{\left(R_a \to R_\beta R_\gamma\right) \in G} \left(f(\mu, s, \beta) \wedge f(\ell - \mu, s + \mu, \gamma) \right)$$

7. **Output:** $w \in L(G) \iff f(n, 1, 1) = 1$.

Analysis

Assume $G = \{R_1, R_2, \ldots, R_r\}$ with start symbol R_1

- $f\left(\text{length}, \text{start pos}, \text{variable index}\right)$.
- Number of subproblems: $O(rn^2)$
- Space: $O(rn^2)$
- Time to evaluate a subproblem from previous ones: $O(|P|n)$
  P is set of rules
- Total time: $O(|P|rn^3)$ which is polynomial in both $|w|$ and $|G|$. For fixed G the run time is cubic in input string length.
- Running time can be improved to $O(n^3|P|)$.
- Not practical for most programming languages. Most languages assume restricted forms of CFGs that enable more efficient parsing algorithms.

Analysis

Assume $G = \{R_1, R_2, \ldots, R_r\}$ with start symbol R_1

- $f\left(\text{length, start pos, variable index}\right)$.
- Number of subproblems: $O(rn^2)$
- Space: $O(rn^2)$
- Time to evaluate a subproblem from previous ones: $O(|P|n)$
  P is set of rules
- Total time: $O(|P|rn^3)$ which is polynomial in both $|w|$ and $|G|$. For fixed G the run time is cubic in input string length.
- Running time can be improved to $O(n^3|P|)$.
- Not practical for most programming languages. Most languages assume restricted forms of CFGs that enable more efficient parsing algorithms.

Analysis

Assume $G = \{R_1, R_2, \ldots, R_r\}$ with start symbol R_1

- $f\left(\text{length}, \text{start pos}, \text{variable index}\right)$.
- Number of subproblems: $O(rn^2)$
- Space: $O(rn^2)$
- Time to evaluate a subproblem from previous ones: $O(|P|n)$
  P is set of rules
- Total time: $O(|P|rn^3)$ which is polynomial in both $|w|$ and $|G|$. For fixed G the run time is cubic in input string length.
- Running time can be improved to $O(n^3|P|)$.
- Not practical for most programming languages. Most languages assume restricted forms of CFGs that enable more efficient parsing algorithms.

CYK Algorithm

```
Input string:  X = x_1 ... x_n.
Input grammar G:  r nonterminal symbols R_1...R_r,  R_1 start symbol.
```

$P[n][n][r]$: Array of booleans. Initialize all to FALSE

for $s = 1$ to n **do**
    for each unit production $R_v \to x_s$ **do**
        $P[1][s][v] \leftarrow$ TRUE
for $\ell = 2$ to n **do** // Length of span
    for $s = 1$ to $n - \ell + 1$ **do** // Start of span
        for $\mu = 1$ to $\ell - 1$ **do** // Partition of span
            for all $(R_a \to R_\beta R_\gamma) \in G$ **do**
                if $P[p][s][\beta]$ and $P[\ell - \mu][s + \mu][\gamma]$ **then**
                    $P[\ell][s][a] \leftarrow$ TRUE
if $P[n][1][1]$ is TRUE **then**
    return ''X is member of language''
else
    return ''X is not member of language''