Advanced Algorithmics (6EAP)
Dynamic Programming

Jaak Vilo
2011 Spring
Example

• Wind has blown away the +, *, (, ) signs
• What’s the maximal value?
• Minimal?

2  1  7  1  4  3
• 2 1 7 1 4 3

• $(2+1)*7*(1+4)*3 = 21*15 = 315$

• $2*1 + 7 + 1*4 + 3 = 17$
• Q: How to maximize the value of any expression?

\[
2 4 5 1 9 8 12 1 9 8 7 2 4 4 1 1 2 3 = ?
\]
Dynamic programming, like the divide-and-conquer method, solves problems by combining the solutions to subproblems.

- Dünaamiline planeerimine.

Divide-and-conquer algorithms partition the problem into independent subproblems, solve the subproblems recursively, and then combine their solutions to solve the original problem.

In contrast, dynamic programming is applicable when the subproblems are not independent, that is, when subproblems share subsubproblems.
Dynamic programming

- Avoid calculating repeating subproblems

- \( \text{fib}(1) = \text{fib}(0) = 1 \);
- \( \text{fib}(n) = \text{fib}(n-1) + \text{fib}(n-2) \)

- Although natural to encode (and a useful task for novice programmers to learn about recursion) recursively, this is inefficient.
Structure within the problem

• The fact that it is not a tree indicates overlapping subproblems.
• A dynamic-programming algorithm **solves every subsubproblem just once** and then saves its answer in a table, thereby avoiding the work of recomputing the answer every time the subsubproblem is encountered.
Topp-down (recursive, memoized)

• **Top-down approach**: This is the direct fall-out of the recursive formulation of any problem. If the solution to any problem can be formulated recursively using the solution to its subproblems, and if its subproblems are overlapping, then one can easily memoize or store the solutions to the subproblems in a table. Whenever we attempt to solve a new subproblem, we first check the table to see if it is already solved. If a solution has been recorded, we can use it directly, otherwise we solve the subproblem and add its solution to the table.
Bottom-up

- **Bottom-up approach**: This is the more interesting case. Once we formulate the solution to a problem recursively as in terms of its subproblems, we can try reformulating the problem in a bottom-up fashion: try solving the subproblems first and use their solutions to build-on and arrive at solutions to bigger subproblems. This is also usually done in a tabular form by iteratively generating solutions to bigger and bigger subproblems by using the solutions to small subproblems. For example, if we already know the values of $F_{41}$ and $F_{40}$, we can directly calculate the value of $F_{42}$. 
Dynamic programming is typically applied to *optimization problems*. In such problems there can be many possible solutions. Each solution has a value, and we wish to find a solution with the optimal (minimum or maximum) value.

We call such a solution *an* optimal solution to the problem, as opposed to *the* optimal solution, since there may be several solutions that achieve the optimal value.
The development of a dynamic-programming algorithm can be broken into a sequence of four steps.

1. Characterize the structure of an optimal solution.
2. Recursively define the value of an optimal solution.
3. Compute the value of an optimal solution in a bottom-up fashion.
4. Construct an optimal solution from computed information.
Matrix multiplication

- for $i=1..n$
  - for $j = 1..k$
    - $c_{ij} = \sum_{x=1..m} a_{ix} b_{xj}$

$O(nmk)$
MATRIX-MULTIPLY($A,B$)

1 if columns [$A$] ≠ rows [$B$]
2 then error "incompatible dimensions"
3 else for $i = 1$ to rows [$A$]
4 do for $j = 1$ to columns [$B$]
5 do $C[i,j] = 0$
6 for $k = 1$ to columns [$A$]
7 do $C[i,j] = C[i,j] + A[i,k] \times B[k,j]$
8 return $C$
184

**Figure 6.6** $A \times B \times C \times D = (A \times (B \times C)) \times D$.

(a) $A \times (B \times C) \times D$
- $A$ is a $50 \times 20$ matrix.
- $B$ is a $20 \times 1$ matrix.
- $C$ is a $1 \times 10$ matrix.
- $D$ is a $10 \times 100$ matrix.

(b) $A \times (B \times C) \times D$
- $A$ is a $50 \times 20$ matrix.
- $B \times C$ is a $20 \times 10$ matrix.
- $D$ is a $10 \times 100$ matrix.

(c) $A \times (B \times C) \times D$
- $A \times (B \times C)$ is a $50 \times 10$ matrix.
- $D$ is a $10 \times 100$ matrix.

(d) $(A \times (B \times C)) \times D$
- $(A \times (B \times C))$ is a $50 \times 100$ matrix.
- $D$ is a $10 \times 100$ matrix.

### 6.5 Chain matrix multiplication
Multiplying an \( m \times n \) matrix by an \( n \times p \) matrix takes \( mnp \) multiplications, to a good enough approximation. Using this formula, let’s compare several different ways of evaluating \( A \times B \times C \times D \):

<table>
<thead>
<tr>
<th>Parenthesization</th>
<th>Cost computation</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A \times ((B \times C) \times D) )</td>
<td>( 20 \cdot 1 \cdot 10 + 20 \cdot 10 \cdot 100 + 50 \cdot 20 \cdot 100 )</td>
<td>120,200</td>
</tr>
<tr>
<td>( (A \times (B \times C)) \times D )</td>
<td>( 20 \cdot 1 \cdot 10 + 50 \cdot 20 \cdot 10 + 50 \cdot 10 \cdot 100 )</td>
<td>60,200</td>
</tr>
<tr>
<td>( (A \times B) \times (C \times D) )</td>
<td>( 50 \cdot 20 \cdot 1 + 1 \cdot 10 \cdot 100 + 50 \cdot 1 \cdot 100 )</td>
<td>7,000</td>
</tr>
</tbody>
</table>

As you can see, the order of multiplications makes a big difference in the final running time! Moreover, the natural greedy approach, to always perform the cheapest matrix multiplication available, leads to the second parenthesization shown here and is therefore a failure.
### Parenthesization

<table>
<thead>
<tr>
<th>Parenthesization</th>
<th>Cost computation</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \times ((B \times C') \times D)$</td>
<td>$20 \cdot 1 \cdot 10 + 20 \cdot 10 \cdot 100 + 50 \cdot 20 \cdot 100$</td>
<td>120,200</td>
</tr>
<tr>
<td>$(A \times (B \times C')) \times D$</td>
<td>$20 \cdot 1 \cdot 10 + 50 \cdot 20 \cdot 10 + 50 \cdot 10 \cdot 100$</td>
<td>60,200</td>
</tr>
<tr>
<td>$(A \times B) \times (C \times D)$</td>
<td>$50 \cdot 20 \cdot 1 + 1 \cdot 10 \cdot 100 + 50 \cdot 1 \cdot 100$</td>
<td>7,000</td>
</tr>
</tbody>
</table>
• The matrix-chain multiplication problem can be stated as follows: given a chain <A₁, A₂, . . . , Aₙ> of n matrices
• matrix Aᵢ has dimension pᵢ₋₁ X pᵢ
• fully parenthesize the product A₁ A₂...Aₙ in a way that minimizes the number of scalar multiplications.
$A_1A_2A_3A_4$

- $(A_1(A_2(A_3A_4)))$,
- $(A_1((A_2A_3)A_4))$,
- $((A_1A_2)(A_3A_4))$,
- $(A_1(A_2A_3))A_4$,
- $((A_1A_2)A_3)A_4$,
• Denote the number of alternative parenthesizations of a sequence of $n$ matrices by $P(n)$.

• Since we can split a sequence of $n$ matrices between the $k$th and $(k + 1)$st matrices for any $k = 1, 2, \ldots, n - 1$ and then parenthesize the two resulting subsequences independently, we obtain the recurrence

\[
P(n) = \begin{cases} 
1 & \text{if } n = 1, \\
\sum_{k=1}^{n-1} P(k)P(n-k) & \text{if } n \geq 2.
\end{cases}
\]
Problem 13-4 asked you to show that the solution to this recurrence is the sequence of *Catalan numbers*:

\[ C(n) = \frac{1}{n+1} \binom{2n}{n} \]

\[ = \Omega(\frac{4^n}{\sqrt{n^3}}) \]

- \( P(n) = C(n - 1) \), where the number of solutions is thus exponential in \( n \), and the brute-force method of exhaustive search is therefore a poor strategy for determining the optimal parenthesization of a matrix chain.
Let’s crack the problem

\[ A_{i..j} = A_i \cdot A_{i+1} \cdot \cdots \cdot A_j \]

- Optimal parenthesization of \( A_1 \cdot A_2 \cdot \cdots \cdot A_n \) splits at some \( k, k+1 \).
- Optimal = \( A_{1..k} \cdot A_{k+1..n} \)

\[ T(A_{1..n}) = T(A_{1..k}) + T(A_{k+1..n}) + T(A_{1..k} \cdot A_{k+1..n}) \]

- \( T(A_{1..k}) \) must be optimal for \( A_1 \cdot A_2 \cdot \cdots \cdot A_k \)
Recursion

• $m[i, j]$ - minimum number of scalar multiplications needed to compute the matrix $A_{i..j}$;

• $m[i,i] = 0$

• $cost(A_{i..k} \cdot A_{k+1..j}) = p_{i-1} p_k p_j$

• $m[i, j] = m[i, k] + m[k + 1, j] + p_{i-1} p_k p_j$.

• This recursive equation assumes that we know the value of $k$, which we don't. There are only $j - i$ possible values for $k$, however, namely $k = i, i + 1, \ldots, j - 1$. 
• Since the optimal parenthesesization must use one of these values for $k$, we need only check them all to find the best. Thus, our recursive definition for the minimum cost of parenthesesizing the product $A_i A_{i+1} \ldots A_j$ becomes

$$m[i, j] = \begin{cases} 0 & \text{if } i = j, \\ \min_{i \leq k < j} \{ m[i, k] + m[k + 1, j] + p_{i-1} p_k p_j \} & \text{if } i < j. \end{cases}$$ (16.2)

• To help us keep track of how to construct an optimal solution, let us define $s[i, j]$ to be a value of $k$ at which we can split the product $A_i A_{i+1} \ldots A_j$ to obtain an optimal parenthesesization. That is, $s[i, j]$ equals a value $k$ such that $m[i, j] = m[i, k] + m[k + 1, j] + p_{i-1} p_k p_j$. 
Recursion

• Checks all possibilities...

• But – there is only a few subproblems – choose \(i, j\) s.t. \(1 \leq i \leq j \leq n\) – \(O(n^2)\)

• A recursive algorithm may encounter each subproblem many times in different branches of its recursion tree. This property of overlapping subproblems is the second hallmark of the applicability of dynamic programming.
MATRIX-CHAIN-ORDER(p)

1 \ n \leftarrow \text{length}[p] - 1
2 \ \textbf{for} \ i \leftarrow 1 \ \textbf{to} \ n \\
3 \ \quad \textbf{do} \ m[i, i] \leftarrow 0 \\
4 \ \textbf{for} \ l \leftarrow 2 \ \textbf{to} \ n \\
5 \ \quad \textbf{for} \ i \leftarrow 1 \ \textbf{to} \ n - l + 1 \\
6 \ \quad \quad \textbf{do} \ j \leftarrow i + l - 1 \\
7 \ \quad \quad m[i, j] \leftarrow \infty \\
8 \ \quad \textbf{for} \ k \leftarrow i \ \textbf{to} \ j - 1 \\
9 \ \quad \quad \textbf{do} \ q \leftarrow m[i, k] + m[k + 1, j] + p_{i-1}p_kp_j \\
10 \ \quad \quad \textbf{if} \ q < m[i, j] \\
11 \ \quad \quad \quad \textbf{then} \ m[i, j] \leftarrow q \quad \quad \quad \textbf{// new best value q} \\
12 \ \quad \quad s[i, j] \leftarrow k \quad \quad \quad \textbf{// achieved at mid point k} \\
13 \ \textbf{return} \ m \ \text{and} \ s
Example

$$((A_1(A_2A_3))((A_4A_5)A_6))$$

Matrix dimensions:

- $A_1$ $30 \times 35$
- $A_2$ $35 \times 15$
- $A_3$ $15 \times 5$
- $A_4$ $5 \times 10$
- $A_5$ $10 \times 20$
- $A_6$ $20 \times 25$

$$m_{2, 5} = \min \left\{ \begin{array}{l}
m[2, 2] + m[3, 5] + p_1 p_2 p_3 = 0 + 2500 + 35 \cdot 15 \cdot 20 = 13000, \\
m[2, 3] + m[4, 5] + p_1 p_3 p_5 = 2625 + 1000 + 35 \cdot 5 \cdot 20 = 7125,
\end{array} \right.$$
A simple inspection of the nested loop structure of MATRIX-CHAIN-ORDER yields a running time of $O(n^3)$ for the algorithm. The loops are nested three deep, and each loop index ($l$, $i$, and $k$) takes on at most $n$ values.

- Time $\Omega(n^3) \Rightarrow \Theta(n^3)$
- Space $\Theta(n^2)$
• Step 4 of the dynamic-programming paradigm is to construct an optimal solution from computed information.

• Use the table $s[1 \ldots n, 1 \ldots n]$ to determine the best way to multiply the matrices.
Multiply using S table

MATRIX-CHAIN-MULTIPLY(A, s, i, j)
1 if \( j > i \)
2 then \( X = \text{MATRIX-CHAIN-MULTIPLY}(A, s, i, s[i, j]) \)
3 \( Y = \text{MATRIX-CHAIN-MULTIPLY}(A, s, s[i, j]+1, j) \)
4 return \( \text{MATRIX-MULTIPLY}(X, Y) \)
5 else return \( A_i \)

((A_1(A_2A_3))((A_4A_5)A_6))
Elements of dynamic programming

• **Optimal substructure** within an optimal solution

• Overlapping subproblems

• Memoization
• A memoized recursive algorithm maintains an entry in a table for the solution to each subproblem. Each table entry initially contains a special value to indicate that the entry has yet to be filled in. When the subproblem is first encountered during the execution of the recursive algorithm, its solution is computed and then stored in the table. Each subsequent time that the subproblem is encountered, the value stored in the table is simply looked up and returned. (tabulated)

• This approach presupposes that the set of all possible subproblem parameters is known and that the relation between table positions and subproblems is established. Another approach is to memoize by using hashing with the subproblem parameters as keys.
Overlapping subproblems
Longest Common Subsequence (LCS)
Optimal triangulation

The problem is to find a triangulation that minimizes the sum of the weights of the triangles in the triangulation.

Two ways of triangulating a convex polygon. Every triangulation of this 7-sided polygon has $7 - 3 = 4$ chords and divides the polygon into $7 - 2 = 5$ triangles.
Parse trees. (a) The parse tree for the parenthesized product \(((A_1 (A_2 A_3))(A_4 (A_5 A_6)))\) and for the triangulation of the 7-sided polygon (b) The triangulation of the polygon with the parse tree overlaid. Each matrix \(A_i\) corresponds to the side \(v_{i-1} v_i\) for \(i = 1, 2, \ldots, 6\).
Optimal triangulation

\[ t[i, j] = \begin{cases} 0 & \text{if } i = j, \\ \min_{i \leq k \leq j-1} \{ t[i, k] + t[k + 1, j] + w(\Delta v_{i-1} v_k v_j) \} & \text{if } i < j \end{cases} \] (16.7)
Text Algorithms (6EAP)

Similarity measures

Jaak Vilo
2010 fall
Similarity

• How can we measure the similarity of two strings?

• When are the two things “almost” the same?
Edit distance (Levenshtein distance)

- Smallest nr of edit operations to convert one string into the other
• **Definition** The edit distance \( D(A,B) \) between strings \( A \) and \( B \) is the minimal number of edit operations to change \( A \) into \( B \). Allowed edit operations are deletion of a single letter, insertion of a letter, or replacing one letter with another.

• Let \( A = a_1 \ a_2 \ \ldots \ a_m \) and \( B = b_1 \ b_2 \ \ldots \ b_m \).
  
  – E1: **Deletion** \( a_i \rightarrow \varepsilon \)
  
  – E2: **Insertion** \( \varepsilon \rightarrow b_i \)
  
  – E3: **Substitution** \( a_i \rightarrow b_j \) (if \( a_i \neq b_j \))

• Other possible variants:
  
  – E4: **Transposition** \( a_ia_{i+1} \rightarrow b_jb_{j+1} \) and \( a_i=b_{j+1} \) ja \( a_{i+1}=b_j \)
    
    (e.g. lecture \( \rightarrow \) letcure)
How can we calculate this?

\[
\begin{array}{|c|c|}
\hline
\alpha & a \\
\hline
\beta & b \\
\hline
\end{array}
\]

\[
\text{D}(\alpha a, \beta b) =
\begin{cases}
1. & \text{D}(\alpha, \beta) & \text{if } a=b \\
2. & \text{D}(\alpha, \beta)+1 & \text{if } a \neq b \\
3. & \text{D}(\alpha a, \beta)+1 \\
4. & \text{D}(\alpha, \beta b)+1 \\
\end{cases}
\]

\[
\text{min}
\]
How can we calculate this efficiently?

\[ D(S,T) = \min \begin{cases} 
1. & D(S[1..n-1], T[1..m-1]) + (S[n]=T[m])? 0 : 1 \\
2. & D(S[1..n], T[1..m-1]) +1 \\
3. & D(S[1..n-1], T[1..m]) +1 
\end{cases} \]

Define: \( d(i,j) = D(S[1..i], T[1..j]) \)

\[ d(i,j) = \min \begin{cases} 
1. & d(i-1,j-1) + (S[n]=T[m])? 0 : 1 \\
2. & d(i, j-1) +1 \\
3. & d(i-1, j) +1 
\end{cases} \]
Recursion?
Recursion?
Algorithm Edit distance D(A,B) using Dynamic Programming (DP)

Input: A=a₁a₂...aₙ, B=b₁b₂...bₘ
Output: Value $d_{mn}$ in matrix $(d_{ij})$, $0 \leq i \leq m$, $0 \leq j \leq n$.

for $i=0$ to $m$ do $d_{i0}=i$ ;

for $j=0$ to $n$ do $d_{0j}=j$ ;

for $j=1$ to $n$ do

  for $i=1$ to $m$ do

    $d_{ij} = \min( d_{i-1,j-1} + (\text{if } a_i == b_j \text{ then } 0 \text{ else } 1),$
    $\quad d_{i-1,j} + 1, \quad d_{i,j-1} + 1 )$

return $d_{mn}$
Dynamic Programming

\[
d(i,j) = \begin{cases} 
0 & \text{if } i = 0 \text{ or } j = 0 \\
d(i-1,j-1) + \begin{cases} 
-1 & \text{if } x_i \neq y_j \\
0 & \text{otherwise} 
\end{cases} & \text{otherwise}
\end{cases}
\]
<table>
<thead>
<tr>
<th>A</th>
<th>b</th>
<th>a</th>
<th>c</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>c</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>b</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>S=</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>T=</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>C</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Edit distance is a metric

• It can be shown, that $D(A,B)$ is a metric
  – $D(A,B) \geq 0$, $D(A,B)=0$ iff $A=B$
  – $D(A,B) = D(B,A)$
  – $D(A,C) \leq D(A,B) + D(B,C)$
Alignment

indust-r-y-
in---terest
Path of edit operations

• Optimal solution can be calculated afterwards
  – Quite typical in dynamic programming

\[
\begin{align*}
  d_{i-1,j-1} & \quad d_{i-1,j} \\
  d_{i,j-1} & \quad d_{i,j}
\end{align*}
\]

• Memorize sets \texttt{pred[i,j]} depending from where the \( d_{ij} \) was reached.
Three possible minimizing paths

- Add into pred[i,j]
  - (i-1,j-1) if \( d_{ij} = d_{i-1,j-1} + (\text{if } a_i == b_j \text{ then } 0 \text{ else } 1) \)
  - (i-1,j) if \( d_{ij} = d_{i-1,j} + 1 \)
  - (i,j-1) if \( d_{ij} = d_{i,j-1} + 1 \)
The path (in reverse order) $\epsilon \rightarrow c_6, b_5 \rightarrow b_5, c_4 \rightarrow c_4, a_3 \rightarrow a_3, a_2 \rightarrow b_2, b_1 \rightarrow a_1$. 
Multiple paths possible

• All paths are correct

• There can be many (how many?) paths
What are the other questions in edit distance calculations?

- Space complexity
- Time complexity
- Other ways to look at the algorithm(s)
- Applications
- More complex notions of similarity
- ...
Space can be reduced

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>a</th>
<th>c</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a b</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>b a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>a b</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>b c</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>c b</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

$C^{[m]} = \begin{bmatrix} 3 & 2 & 3 \\ 2 & 1 & 3 \\ 2 & 1 & 3 \end{bmatrix}$
Calculation of D(A,B) in space $\Theta(m)$

**Input:** $A=a_1a_2...a_m$, $B=b_1b_2...b_n$ (choose $m<=n$)

**Output:** $d_{mn} = D(A,B)$

for $i=0$ to $m$ do $C[i]=i$

for $j=1$ to $n$ do
    $C = C[0]$; $C[0]=j$; 
    for $i=1$ to $m$ do
        $d = \min( C + (if \ a_i==b_j \ then \ 0 \ else \ 1), \ C[i-1] + 1, \ C[i] + 1 )$
        $C = C[i]$  // memorize new “diagonal” value
        $C[i] = d$

write $C[m]$

Time complexity is $\Theta(mn)$ since $C[0..m]$ is filled n times
Shortest path in the graph

Shortest path in the graph


All nodes at distance 1 from source
|   | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
| 1 |   | A | B | C | D | E | F | G | H | I | J | K | L |   | A | C | A | C | C | C |
| 2 | T | A | A | A | C | B | C | B | A | B | B | A | A | C | A | C | C | C | B |
| 3 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 7 | A |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 8 | A | 2 | 1 | 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 9 | A | 3 | 2 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 10| C | 4 | 3 | 3 | 2 | 3 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 11| B | 5 | 4 | 3 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 12| B | 6 | 5 | 4 | 4 | 3 | 2 | 2 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 13| C | 7 | 6 | 5 | 5 | 4 | 3 | 2 | 3 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 14| A | 8 | 7 | 6 | 5 | 5 | 4 | 3 | 2 | 3 | 3 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 15| C | 9 | 8 | 7 | 6 | 5 | 4 | 4 | 3 | 4 | 3 | 4 | 4 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 16| A | 10| 9 | 9 | 8 | 7 | 6 | 5 | 5 | 4 | 4 | 4 | 5 | 6 | 7 | 7 | 8 | 9 | 10 | 11 |
| 17| B | 11| 10| 9 | 8 | 7 | 6 | 6 | 5 | 5 | 4 | 5 | 4 | 5 | 6 | 7 | 8 | 8 | 9 | 10 |
| 18| B | 12| 11| 10| 9 | 8 | 7 | 7 | 6 | 6 | 5 | 5 | 4 | 5 | 4 | 5 | 6 | 7 | 8 | 9 |
| 19| A | 13| 12| 11| 10| 9 | 8 | 7 | 6 | 7 | 6 | 5 | 5 | 4 | 5 | 4 | 5 | 6 | 7 | 8 |
| 20| A | 14| 13| 12| 11| 10| 9 | 8 | 8 | 7 | 6 | 6 | 6 | 5 | 4 | 5 | 6 | 7 | 8 |
| 21| C | 15| 14| 13| 12| 11| 10| 9 | 9 | 8 | 8 | 7 | 7 | 6 | 5 | 4 | 5 | 6 | 7 |
| 22| A | 16| 15| 14| 13| 12| 11| 10| 9 | 9 | 8 | 8 | 7 | 6 | 6 | 5 | 4 | 5 | 6 |
| 23| C | 17| 16| 15| 14| 13| 12| 11| 11| 10| 9 | 9 | 8 | 8 | 7 | 6 | 5 | 4 | 5 |
| 24| A | 18| 17| 16| 15| 14| 13| 12| 12| 11| 10| 10| 9 | 8 | 7 | 6 | 5 | 4 | 5 |
| 25| B | 19| 18| 17| 16| 15| 14| 13| 12| 12| 11| 10| 10| 9 | 8 | 7 | 6 | 5 | 5 |
Observations?

• Shortest path is close to the diagonal
  – If a short distance path exists

• Values along any diagonal can only increase (by at most 1)
Diagonal

Property of any diagonal: The values of matrix \((d_{ij})\) can on any specific diagonal either increase by 1 or stay the same.

Diagonal nr. 2, \(d_{02}, d_{13}, d_{24}, d_{35}, d_{46}\)

Diagonal \(k\), \(-m \leq k \leq n\),
s.t. diagonal \(k\) contains only \(d_{ij}\) where \(j-i = k\).
Diagonal lemma

**Lemma:** For each $d_{ij}$, $1 \leq i \leq m$, $1 \leq j \leq n$ holds: $d_{ij} = d_{i-1,j-1}$ or $d_{ij} = d_{i-1,j-1} + 1$. (notice that $d_{ij}$ and $d_{i-1,j-1}$ are on the same diagonal)

**Proof:** Since $d_{ij}$ is an integer, show:

- $d_{ij} \leq d_{i-1,j-1} + 1$
- $d_{ij} \geq d_{i-1,j-1}$

From the definition of edit distance $1$. holds since $d_{ij} \leq d_{i-1,j-1} + 1$

Induction on $i+j$:

- Basis is trivial when $i=0$ or $j=0$ (if we agree that $d_{-1,j-1}=d_{0}$)
- Induction step: there are 3 possibilities -
  - On minimization the $d_{ij}$ is calculated from entry $d_{i-1,j-1}$, or $d_{ij} \geq d_{i-1,j-1}$
  - On minimization the $d_{ij}$ is calculated from entry $d_{i-1,j}$, or $d_{ij} = d_{i-1,j} + 1 \geq d_{i-2,j-1} + 1 \geq d_{i-1,j-1}$
  - On minimization the $d_{ij}$ is calculated from entry $d_{i,j-1}$. Analogical to 2.
- Hence, $d_{i-1,j-1} \leq d_{ij}$
Transform the matrix into $f_{kp}$

- For each diagonal only show the position (row index) where the value is increased by 1.
- Also, one can restrict the matrix ($d_{ij}$) to only this part where $d_{ij} \leq d_{mn}$ since only those $d_{ij}$ can be on the shortest path.
- We'll use the matrix ($f_{kp}$) that represents the diagonals of $d_{ij}$
  - $f_{kp}$ is a row index $i$ from $d_{ij}$, such that on diagonal $k$ the value $p$ reaches row $i$ ($d_{ij}=p$ and $j-i=k$).
  - Initialization: $f_{0,-1}=-1$ and $f_{kp}=-\infty$ when $p \leq |k|-1$ ;
  - $d_{mn}=p$, such that $f_{n-m,p}=m$
Example (compare to prev. ex)

- In orig. matrix \((d_{ij})\) there are 42 values, and in diagonal matrix \((f_{kp})\) 22 values (excl. \(-\infty\)).

- **Note:** The values in \((f_{kp})\) are not comfortably organized. With appropriate index transformations it is possible to give the \((f_{kp})\) more regular shape.
Calculating matrix \( (f_{kp}) \) by columns

- Assume the column \( p-1 \) has been calculated in \( (f_{kp}) \), and we want to calculate \( f_{kp} \). (the region of \( d_{ij}=p \))
- On diagonal \( k \) values \( p \) reach at least the row \( t=\max \left( f_{k,p-1}+1, f_{k-1,p-1}, f_{k+1,p-1}+1 \right) \) if the diagonal \( k \) reaches so far.
- If on row \( t+1 \) additionally \( a_i = b_j \) on the same diagonal, then \( d_{ij} \) cannot increase, and value \( p \) reaches row \( t+1 \).
- Repeat previous step until \( a_i \neq b_j \) on diagonal \( k \).
• $f_{k,p-1} + 1$ - same diagonal
• $f_{k-1,p-1}$ - diagonal below
• $f_{k+1,p-1} + 1$ - diagonal above
Algorithm A(): calculate $f_{kp}$

A(k,p)
1. $t = \max( f_{k,p-1} +1 , f_{k-1,p-1} , f_{k+1,p-1} +1)$
2. while $a_{t+1} == b_{t+1+k}$ do $t = t+1$
3. $f_{kp} = \text{if } t>m \text{ or } t+k >n \text{ then undefined } \text{ else } t$
Example (compare to prev. ex)

Diagonals x edit distances

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>b</td>
<td>-5</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>-4</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>-3</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>-2</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td>-∞</td>
<td>-∞</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-∞</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-∞</td>
<td>-∞</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>-∞</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ f(0,2) = \max(3,2,3) = 2 \]

\[ f(1,2) \]

Algorithm: Diagonal method by columns

\[ p = -1 \]

**while** \( f_{n-m,p} \neq m \)

\[ p = p + 1 \]

**for** \( k = -p \) to \( p \) **do**  

// \( f_{kp} = A(k,p) \)

\[ t = \max( f_{k,p-1} + 1, f_{k-1,p-1}, f_{k+1,p-1} + 1 ) \]

**while** \( a_{t+1} = b_{t+1+k} \) **do**  

\[ t = t + 1 \]

\( f_{kp} = \) **if** \( t > m \) **or** \( t + k > n \) **then** undefined **else** \( t \)
• p can only occur on diagonals \(-p \leq k \leq p\).
• Method can be improved since \(k\) is often such that \(f_{kp}\) is undefined.
• We can decrease values of \(k\):
  – \(-m \leq k \leq n\) (diagonal numbers)
  – Let \(m \leq n\) and \(d_{ij}\) on diagonal \(k\).
    • if \(-m \leq k \leq 0\) then \(|k| \leq d_{ij} \leq m\)
    • if \(1 \leq k \leq n\) then \(k \leq d_{ij} \leq k+m\)
    • Hence, \(-m \leq k \leq m\) if \(p \leq m\) and \(p-m \leq k \leq p\) if \(p \geq m\)
Calculating matrix \((f_{kp})\) by columns (3)

Algorithm: Diagonaal meetod

Input: \(A = a_1a_2...a_m, \ B = b_1b_2...b_n, \ m \leq n\) (replace \(A, B\) if needed)

Output: \(p = D(A, B)\) and matrix \((f_{kp})\) (the part that has been calculated)

1. \(p = -1; \ r = p - m\)
2. while \(f_{n-m, p} \neq m\)
3. \(p = p + 1; \ r = r + 1;\)
4. if \(r \leq 0\) then \{ \(\text{mini} = -p; \ \text{maxi} = p\) \}
5. else \{ \(\text{mini} = r; \ \text{maxi} = \min(p, n)\) \}
6. for \(k = \text{mini} \) to \(\text{maxi} \) do begin
7. \(t = \max(f_{k, p - 1} + 1, f_{k-1, p - 1}, f_{k+1, p - 1} + 1)\)
8. while \(a_{t+1} = b_{t+1+k}\) do \(t = t + 1\)
9. \(f_{kp} = t\)
10. write \(p\)

- If on line 7 the value \(f_{kp}\) is not yet calculated, then \(f_{kp} = -\infty\)
Some notes

• In applications small $D(A,B)$ are most interesting.
• Can modify the algorithm by providing maximum $t$.
• Hence, $O(tm)$ - the smaller the $t$, the faster the algorithm.
• Space can be reduced by keeping only previous column.
• How to output the shortest path?
  – Relatively simple, in time $O(s)$, outputs a single path.
Extensions to basic edit distance

• New operations

• Variable costs

• ...
Transposition (ab → ba)

• **E4: Transposition**

  \[ a_ia_{i+1} \rightarrow bjb_{j+1}, \text{ s.t. } a_i=b_{j+1} \text{ and } a_{i+1}=b_j \]

• (e.g.: lecture → letcure)

\[
d(i,j) = \min \begin{cases} 
1. & d(i-1,j-1) + (S[n]=T[m])? 0 : 1 \\
2. & d(i, j-1) +1 \\
3. & d(i-1, j) +1 \\
4. & d(i-2,j-2) + ( \text{ if } S[i-1,i] = T[j,j-1] \text{ then } 1 \text{ else } \infty ) 
\end{cases}
\]
• The edit distance algorithm can be changed easily
• Space efficiency can also be achieved using 2 last columns, hence still $O(m)$
• Diagonal method can be modified, since the diagonal lemma holds.
• Algorithms can be modified in a relatively straightforward manner
Longest common subsequences

• **Definition.** String $C=c_1c_2...c_r$ is a subsequence (alamjada) of $A=a_1a_2...a_m$ if by removing from $A$ null or more characters, one can get $C$.

• String $C=c_1c_2...c_r$ is the **longest common subsequence**, LCS (pikim ühine alamjada) of $A=a_1a_2...a_m$ and $B=b_1b_2...b_n$ if $C$ is the longest string that is both the subsequence of $A$ and $B$.

• $C=LCS(A,B)$

• The length of $LCS(A,B)$, $|C|$, can be used as the similarity measure for $A$ and $B$.

• $LCS(A,B)$ can be calculated similarly to edit distance
\[ |LCS(A,B)| = \left( |A| + |B| - D'(A,B) \right)/2 \]

• Let \( D'(A,B) \) the edit distance where the only allowed operations are insertion and deletion (no replace).

• Theorem
  
  a) \[ |LCS(A,B)| = \left( |A| + |B| - D'(A,B) \right)/2 \]
  b) Lets have two sets \( D'(A,B) \) with optimal nr of changes:
     1. \( a_{i1} \rightarrow \varepsilon, a_{i2} \rightarrow \varepsilon, \ldots a_{ip} \rightarrow \varepsilon \) deletions from A and
     2. \( \varepsilon \rightarrow b_{j1}, \varepsilon \rightarrow b_{j2}, \ldots \varepsilon \rightarrow b_{jr} \) insertions into B.

Then LCS(A,B)=C can be constructed such that, C is A after deletions of 1. and C is B after deletion of all insertions 2. (insertions in 2. are reversely deletions from B).
Proof b)
• According to construction, C is uniquely defined
  C is a subsequence of A as well as B.
  If C was not the longest, then there would be
    C', |C| < |C'| s.t. C'=LCS(A,B).
• But then D'(A,B) ≤ |A|−|C'|+|B|−|C'| < |A|−|C|+|B|−|C| = D'(A,B), which is a contradiction.
• Hence, C=LCS(A,B).

Proof a)
• According to b) |LCS(A,B)|=|A|−p and |LCS(A,B)|=|B|−r, or 2·|LCS(A,B)|=|A|+|B|−(p+r)=|A|+|B|−D'(A,B).
• **Example.** LCS(england, inglismaa) = ngla. 
  D'(england, inglismaa) = 8, |ngla| = 4 = (7 + 9 - 8) / 2.
• Diagonal lemma holds, but the increase always occurs by two.
• Time complexity O(mn), with diag. method O (min(s, m) s) where s = D'(A, B), m = |A|, n = |B|.
• There exists other algorithms for LCS (e.g. Hunt-Szymanski)
• Unix command **diff** compares files row by row and searches the deviations from the LCS of the two files.
Generalized edit distance

• Use more operations E1...En, and to provide different costs to each.
• **Definition.** Let $$x, y \in \Sigma^*$$. Then every $$x \rightarrow y$$ is an edit operation. Edit operation replaces $$x$$ by $$y$$.
  – If $$A=uxv$$ then after the operation, $$A=uyv$$
• We note by $$w(x \rightarrow y)$$ the cost or weight of the operation.
• Cost may depend on $$x$$ and/or $$y$$. But we assume $$w(x \rightarrow y) \geq 0$$. 
Generalized edit distance

• If operations can only be applied in parallel, i.e. the part already changed cannot be modified again, then we can use the dynamic programming.

• Otherwise it is an algorithmically unsolvable problem, since question - can A be transformed into B using operations of G, is unsolvable.

• The diagonal method in general may not be applicable.

• But, since each diversion from diagonal, the cost slightly increases, one can stay within the narrow region around the diagonal.
Applications of generalized edit distance

• Historic documents, names
• Human language and dialects
• Transliteration rules from one alphabet to another
  e.g. Tõugu => Tyugu (via Russian)
• ...
Examples

Päringuks kulus: ~2 sekund(it);
<table>
<thead>
<tr>
<th>Otsisõna:</th>
<th>tajzhelo</th>
</tr>
</thead>
</table>

Pahun vali sõnadest (0), milles otsing sooritatakse:
- Kohanimed
- Venekeelsed sõnad
- Rossinuse sõnad
- Stahl sõnad
- Mülleri sõnad
- ET-Eesti-Englise
- EN-Englise-Eesti

Milliseid vasteid otsida?
- Täpsed vasted
- Algusosa vasted
- Keskosa vasted
- Lõpuosa vasted

Täiendavad otsinguülimused
- Tööstuslik otsing
Maksimaalne teiseduskangus: 1,5
Vastete maksimaalne arv (ühes tulbas):

### Täpsed vasted

<table>
<thead>
<tr>
<th>Täpsed vasted</th>
<th>Keskosa vasted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TÁJŽHELO</strong></td>
<td></td>
</tr>
<tr>
<td>/Venekeelsed sõnad/</td>
<td>/Venekeelsed sõnad/</td>
</tr>
<tr>
<td>1.100800</td>
<td>0.100600</td>
</tr>
</tbody>
</table>

| TÄJŽHELO       |
| /Venekeelsed sõnad/ | /Venekeelsed sõnad/ |
| 1.100600      | 0.100600       |

| TÄŽHELO       |
| /Venekeelsed sõnad/ | /Venekeelsed sõnad/ |
| 1.100800      | 0.100600       |

| TÁŽHELO       |
| /Venekeelsed sõnad/ | /Venekeelsed sõnad/ |
| 1.100600      | 0.100600       |

| TÁŽHELO       |
| /Venekeelsed sõnad/ | /Venekeelsed sõnad/ |
| 1.100800      | 0.100600       |

| TÁŽHELO       |
| /Venekeelsed sõnad/ | /Venekeelsed sõnad/ |
| 1.100600      | 0.100600       |
### Otsisõna:
- rahvas

### Kohanimed
- Valk Eesti kohanimesid

### Venekeelsed sõnad
- Valk venekeelsed sõnud

### Rossinõuse sõnad
- Sõnad Rossiniuse (~1000 - ~1640) tekstidest

### Stahl sõnad
- Sõnad Stahl (~1600 - ~1657) tekstidest

### Müller sõnad
- Sõnad Müller (~1570 - ~1608) tekstidest

### ET-Eesti-Ingliste
- Eesti keelest vasted Eesti-Ingliste sõnastikust

### EN-Ingliste-Eesti
- Ingliste keelest vasted Ingliste-Eesti sõnastikust

### Millised vasteid otsida?
- Täpsed vasted
- Algusosa vasted
- Keskosa vasted
- Lõpuosa vasted

### Täiendavad otingutingerimused
- Tõstutundlik otsing
  - Maksimaalne
teisenduskangus: 15
  - Vastete
maksimaalne arv: 10
  (ühes tulvas):

### Täpsed vasted
- **Rahvast**
  - 1.084947

### Algusosa vasted
- **Rahvast**
  - 1.084947

### Keskosa vasted
- **Rahvast**
  - 1.084947

### Lõpuosa vasted
- **Rahvast**
  - 1.084947

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829

- **Rahvast**
  - 1.087829
näituseks – näiteks
Ahwrika - Aafrika
weikese - väikese
materjaali - materjali

“kavalam” otsimine
Dush, dušš, dushsh ?
Gorbatšov, Gorbatshov, Горбачов,
Gorbachev
režiim, rezhiim, riim

tuseks -> teks
a -> aa , hw -> f
w -> v , e -> ä
aa -> a
How?

• Apply Aho-Corasick to match for all possible edit operations

• Use minimum over all possible such operations and costs

• Implementation: Reina Käärik, Siim Orasmaa
Possible problems/tasks

• Manually create sensible lists of operations
  – For English, Russian, etc...
  – Old language,

• Improve the speed of the algorithm (testing)

• Train for automatic extraction of edit operations and respective costs from examples of matching words...
Advanced Dynamic Programming

• Robert Giegerich:
  – [http://www.techfak.uni-bielefeld.de/ags/pi/lehre/ADP/](http://www.techfak.uni-bielefeld.de/ags/pi/lehre/ADP/)

• Algebraic dynamic programming
  – Functional style
  – Haskell compiles into C