Advanced Algorithmics (4AP)
Dynamic programming

Jaak Vilo
2009 Spring


• Dynamic programming, like the divide-and-conquer method, solves problems by combining the solutions to subproblems.
  – Dünaamiline programmeerimine/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.

- 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.
• 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 \( j = 1..k \)
    - \( c_{ij} = \sum_{x=1..m} a_{ix} b_{xj} \)

\[
\begin{pmatrix}
A \\
\end{pmatrix}
\begin{pmatrix}
B \\
\end{pmatrix} =
\begin{pmatrix}
C \\
\end{pmatrix}
\]

\( O(nmk) \)

```plaintext
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]
8 return C
```
### 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>$20 \cdot 1 \cdot 10 + 20 \cdot 10 \cdot 100 = 20 \cdot 10 \cdot 100$</td>
<td>$200,000$</td>
</tr>
<tr>
<td>$(A \times (B \times C)) \times D$</td>
<td>$20 \cdot 1 \cdot 10 + 20 \cdot 10 \cdot 100 = 20 \cdot 100 \cdot 100$</td>
<td>$600,000$</td>
</tr>
<tr>
<td>$(A \times B) \times (C \times D)$</td>
<td>$50 \cdot 20 \cdot 1 \cdot 10 + 50 \cdot 1 \cdot 100 = 50 \cdot 1 \cdot 100$</td>
<td>$500,000$</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.
The **matrix-chain multiplication problem** can be stated as follows: given a chain \(<A_1, A_2, \ldots, A_n>\) of \(n\) matrices

- matrix \(A_j\) has dimension \(p_{j-1} \times p_j\)
- fully parenthesize the product \(A_1 A_2 \ldots A_n\) 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**:

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

• \( P(n) = C(n - 1) \), where

\[
P(n) = \Omega\left(\frac{4^n}{n^{3/2}}\right).
\]

• 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 $A_{i..j}$
• $m[i, i] = 0$
• $cost(A_{i..k} \cdot A_{k+1..j}) = p_{i-1}p_kp_j$
• $m[i, j] = m[i, k] + m[k + 1, j] + p_{i-1}p_kp_j$.

• 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 parenthesization 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 parenthesizing the product $A_i A_{i+1} \ldots A_j$ becomes

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

• 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 parenthesization. That is, $s[i, j]$ equals a value $k$ such that $m[i, j] = m[i, k] + m[k + 1, j] + p_{i-1}p_kp_j$. 

Recursion

• Checks all possibilities...

• But – there is only a few subproblems – choose i, j s.t. 1 ≤ i ≤ j ≤ n - O(n²)

• 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(ρ)

```
MATRIX-CHAIN-ORDER(ρ)
1 n = length[ρ] - 1
2 for i = 1 to n
3    do m[i, i] = 0
4 for l = 2 to n
5    do for i = 1 to n - l + 1
6        do j = i + l - 1
7            m[i, j] = ∞
8            for k = i to j - 1
9                do q = m[i, k] + m[k + 1, j] + ρ[i]ρ[k+1]ρ[j]
10                   if q < m[i, j]
11                       then m[i, j] = q
12                       s[i, j] = k
13 return m and s
```
MATRIX-CHAIN-ORDER(p)
1 n ← length[p] - 1
2 for i ← 1 to n
3 do m[i, i] ← 0
4 for l ← 2 to n
5 do for i ← 1 to n - l + 1
6 do j ← i + l - 1
7 m[i, j] ← ∞
8 for k ← i to j - 1
9 do q ← m[i, k] + m[k + 1, j] + p[i-1] · p[j]
10 if q < m[i, j]
11 then m[i, j] ← q
12 s[i, j] ← k
13 return m and s

Example

((A_1(A_2A_3))(A_4A_5A_6))

matrix dimensions:
A_1 30 X 35
A_2 35 X 15
A_3 15 X 5
A_4 5 X 10
A_5 10 X 20
A_6 20 X 25
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

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

\[ ((A_1(A_2A_3))(A_4A_5)A_6)) \]

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.
Longest Common Subsequence (LCS)

<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>i</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

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 
$((A_1(A_2A_3))(A_4(A_5A_6)))$ 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

$$d[i, j] = \begin{cases} 0, & \text{if } i = j, \\ \min_{i \leq k \leq j-1} \{d[i, k] + d[k+1, j] + w(v_{i-1}v_{k}v_{j})\}, & \text{if } i < j. \end{cases} \quad (16.7)$$
Text Algorithms (4AP)
Lecture 5: Similarity measures

Jaak Vilo
2008 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

\[
\begin{array}{c}
\text{INDUSTRY} \\
\text{INTEREST}
\end{array}
\quad \quad
\begin{array}{c}
\text{INDUSTRY} \\
\text{INTEREST}
\end{array}
\]

• 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 \( A = a_1 \ a_2 \ ... \ a_m \) and \( B = b_1 \ b_2 \ ... \ b_m \).
- E1: Deletion \( a_i \rightarrow \epsilon \)
- E2: Insertion \( \epsilon \rightarrow b_i \)
- E3: Substitution \( a_i \rightarrow b_j \) (if \( a_i \neq b_j \))
- Other possible variants:
- E4: Transposition \( a_i a_{i+1} \rightarrow b_j b_{j+1} \) and \( a_i=b_{j+1} \) ja \( a_{i+1}=b_j \) (e.g. lecture \( \rightarrow \) letcure)
How can we calculate this?

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

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

- \( F(0) = 1 \)
- \( F(1) = 1 \)
- \( F(n) = F(n-1) + F(n-2) \)

- 1, 1, 2, 3, 5, 8, ...

```sub fib(int x)
if (x<3) return 1;
else return fib(x-1)+fib(x-2);
```

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 \( (d_{ij}) \), \( 0 \leq i \leq m \), \( 0 \leq j \leq n \).

1. for \( i=0 \) to \( m \) do \( d_{i0} = i \);
2. for \( j=0 \) to \( n \) do \( d_{0j} = j \);
3. for \( j=1 \) to \( n \) do
   1. for \( i=1 \) to \( m \) do
      1. \( d_{ij} = \min( d_{i-1,j-1} + (\text{if } a_i == b_j \text{ then } 0 \text{ else } 1), \)
         \( d_{i-1,j} + 1, \ d_{i,j-1} + 1 ) \)

return \( d_{mn} \)
Dynamic Programming

\[ d(i, j) = \min \{ d(i-1, j-1), d(i-1, j), d(i, j-1) \} + x_{i,j} \]

\[ d(i, j) = \min \{ d(i-1, j-1), d(i-1, j), d(i, j-1) \} + y_{i,j} \]

\[ d(i, j) = \min \{ d(i-1, j-1), d(i-1, j), d(i, j-1) \} + z_{i,j} \]

\[ d(i, j) = \min \{ d(i-1, j-1), d(i-1, j), d(i, j-1) \} + w_{i,j} \]
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

\[d[i-1, j-1] \quad d[i-1, j]\]
\[\downarrow\]
\[d[i, j-1] \quad d[i, j]\]

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

- Add into pred[i,j]
  
  - $d_{i-1,j-1}$ 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) $\epsilon \to c_6, b_5 \to b_5, c_4 \to c_4, a_3 \to a_3, a_2 \to b_2, b_1 \to a_1$. 

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

The path (in reverse order) $\epsilon \to c_6, b_5 \to b_5, c_4 \to c_4, a_3 \to a_3, a_2 \to b_2, b_1 \to a_1$. 


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

Calculation of $D(A,B)$ in space $\Theta(m)$

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

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


All nodes at distance 1 from source
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 \((d_{ij})\) 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 \(d_{ij} = d_{i-1,j} + 1\).

(notice that \(d_{ij}\) and \(d_{i-1,j}\) 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 \(d_{0,1} = d_{0,j}\))
- Induction step: there are 3 possibilities -
  - On minimization the \(d_{ij}\) is calculated from entry \(d_{i-1,j}\) or \(d_{i,j} \geq d_{i-1,j+1}\)
  - On minimization the \(d_{ij}\) is calculated from entry \(d_{i-1,j+1}\) or \(d_{i,j} = d_{i-1,j} + 1\) or \(d_{i,j} + 1 = 12d_{i,j+1} + 12d_{i+1,j+1}\)
  - On minimization the \(d_{ij}\) is calculated from entry \(d_{i+1,j}\). 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 $f_{n-m,p}=m$
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 \(t=\max(f_{k,p-1}+1, f_{k-1,p-1}, f_{k+1,p-1}+1)\) 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. $f_{kp} = \text{if } t > m \text{ or } t+k > n \text{ then undefined else } t$

Example
• Example (compare to prev. ex)

Diagonals x edit distances

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


Algorithm: Diagonal method by columns

\[ p = -1 \]

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

\[ p = p + 1 \]

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

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

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

\[ t = t + 1 \]

\[ f_{kp} = \text{if } t > m \text{ or } t + k > n \text{ then undefined else } t \]

- \( p \) 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 \(p \geq m\)
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
  \[ a_i a_{i+1} \rightarrow b_j b_{j+1}, \text{ s.t. } a_i = b_{j+1} \text{ and } a_{i+1} = b_j \]

• (e.g.: lecture → lecture)

\[
d(i,j) = \min \begin{cases} 
  1. & d(i-1,j-1) + (S[n]=T[m])? 0 : 1 \\
  2. & d(i, j-1) +1 \\
  3. & d(i-1, j) +1 \\
  4. & d(i-2,j-2) + ( \text{ if } S[i-1,i] = T[j-1,j] \text{ then } 1 \text{ else } \infty )
\end{cases}
\]
• 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) \( |LCS(A,B)| = (|A|+|B| - D'(A,B)) / 2 \)
  
  b) Lets have two sets \( D'(A,B) \) with optimal nr of changes:
  
  1. \( \alpha_1 \rightarrow \varepsilon, \alpha_2 \rightarrow \varepsilon, \ldots, \alpha_p \rightarrow \varepsilon \) deletions from \( A \) and
  
  2. \( \varepsilon \rightarrow \beta_1, \varepsilon \rightarrow \beta_2, \ldots, \varepsilon \rightarrow \beta_r \) 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
  
  \( C' < |C'| \) s.t. \( C' = LCS(A,B) \).
  
  - But then \( D'(A,B) \leq |A| - |C'| + |B| - |C'| < |A| - |C'| + |B| - |C| = D'(A,B) \), 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. 
  \[ D'(\text{england, inglismaa})=8, \ |\text{ngla}|=4=(7+9-8)/2. \]

• Diagonal lemma holds, but the increase always occurs by two.

• Time complexity \( O(mn) \), with diag. method 
  \( O(\min(s,m)) \) 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 $E_1...E_n$, 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
näituseks – näiteks
Ahwrika - Aafrika
weikese - väikese
materjaali - materjali

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

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

How?

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

• Use minimum over all possible such operations and costs

• Implementation: Reina Käärik
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...