Dynamic Programming

Some of the slides are courtesy of Charles Leiserson with small changes by Carola Wenk

We look at sequences of characters (strings)
e.g. \( x = "ABCA" \)

**Def:** A subsequence of \( x \) is an sequence obtained from \( x \) by possibly deleting some of its characters (but without changing their order)

**Examples:**
- "ABC"
- "ACA"
- "AA"
- "ABCA"

**Def** A prefix of \( x \), denoted \( x[1..m] \), is the sequence of the first \( m \) characters of \( x \)

**Examples:**
- \( x[1..4] = "ABCA" \)
- \( x[1..3] = "ABC" \)
- \( x[1..2] = "AB" \)
- \( x[1..1] = "A" \)
- \( x[1..0] = "" \)

---

**Longest Common Subsequence (LCS) problem:**
- Given two sequences \( x[1..m] \) and \( y[1..n] \), find a longest subsequence common to them both.
  - (sometimes in the slides we will use \( n \) to denote the lengths of both sequences)

\[ x: A B C B D A B \]
\[ y: B D C A B A \]

Different phrasing: Find a set of a maximum number of segments, such that
- Each segment connects a character of \( x \) to an identical character of \( y \),
- Each character is used at most once
- Segments do not intersect.

\[ \text{BCBA} = \text{LCS}(x, y) \]
Brute-force LCS algorithm

Checking every subsequence of $x$ whether it is also a subsequence of $y$.

Analysis
• Checking $= \Theta(m+n)$ time per subsequence.
• $2^m$ subsequences of $x$

Worst-case running time $= \Theta((m+n)2^m)$
= exponential time.

Towards a better algorithm

Simplification:
1. Look at the length of a longest-common subsequence.
2. Extend the algorithm to find the LCS itself.

Notation: Denote the length of a sequence $s$ by $|s|$.

Strategy: Consider prefixes of $x$ and $y$.
• Define $c[i,j] = |LCS(x[1..i], y[1..j])|$.
• Then, $c[m, n] = |LCS(x, y)|$.

Structure of LCS - recursive formula
Consider a matching length with maximal length ($|LCS(x,y)|$).
Lets look at $x[n]$ (the last character of x). Three cases:
1. $x[m]$ is matched and $x[m] = y[n]$ (e.g., they both contains the same char ‘B’)
2. $x[m]$ matched and $x[m] \neq y[n]$ (e.g., $x[m] = 'B', y[n] = 'C'$)
3. $x[m]$ is not matched
• Lets consider case 1.
• Claim: we could assume that $x[m]$ is matched to $y[n]$. If not, change the matching to include this pair.
• That is, replace the segment $x[m], y[j]$ with the segment $x[m], y[n]$. The new segment cannot cross any previous segments.
• Conclusion: if $x[m] = y[n]$ then $|LCS(m, n)| = 1 + |LCS(m-1, n-1)|$

case 2 and 3 - $x[m] \neq y[n]$
• Observation:
  • It is impossible that
  • $x[m]$ is matched to an element in $y[1..n-1]$ and simultaneously
  • $y[n]$ is matched to an element in $x[1..m-1]$
  • (since it must create a pair of crossing segments).
  • That is, at most one of $x[m], y[n]$ is matched in OPT.
  • Conclusion: Deleting the unmatched character from its sequence will not hurt LCS(x,y).

Conclusion: If $x[m] \neq y[n]$ then

$$|LCS(x,y)| = \max \left\{ |LCS(x[1..m-1], y)|, |LCS(x, y[1..n-1])| \right\}$$

In the example,
$x[1..m-1] = CCB, y[1..n-1] = DBC$
$LCS(x, y) = CC$
Putting it together

- For every $i, j$
- If $x[i] = y[j]$ then $c[i, j] = 1 + c[i - 1, j - 1]$
- If $x[i] \neq y[j]$ then $c[i, j] = \max\{c[i, j - 1], c[i - 1, j]\}$

Dynamic-programming hallmark #1

**Optimal substructure**
An optimal solution to a problem (instance) contains optimal solutions to subproblems.

If $z = \text{LCS}(x, y)$, then any prefix of $z$ is an LCS of a prefix of $x$ and a prefix of $y$.

Recursive algorithm for LCS

$LCS(x, y, i, j) // we\ start\ the\ method\ with\ i=|x|,\ j=|y|$
- if ( $i=0$ or $j=0$) return 0
- if $x[i] = y[j]$
  - then return $LCS(x, y, i-1, j-1) + 1$
- else return $\max\{LCS(x, y, i-1, j), LCS(x, y, i, j-1)\}$

Worst-case: $x[i] \neq y[j]$, for all $i, j$ in which case the algorithm evaluates two subproblems, each with only one parameter decremented.

Recursion tree

$m = 3, n = 4$:

Height = $m + n \Rightarrow$ work potentially $2^{m+n}$ exponential, but we’re solving subproblems already solved!
Dynamic-programming hallmark #2

Overlapping subproblems
A recursive solution contains a “small” number of distinct subproblems repeated many times.

The number of distinct LCS subproblems for two strings of lengths \( m \) and \( n \) is only \( mn \).

Memoization algorithm

**Memoization:** After computing a solution to a subproblem, store it in a table. Subsequent calls check the table to avoid redoing work.

\[
\begin{align*}
\text{LCS}(x, y) & \\
\text{for } i=0 \text{ to } m & \quad c[i, 0] = 0 \\
\text{for } j=0 \text{ to } n & \quad c[0, j] = 0 \\
\text{for } i=1 \text{ to } m & \text{ for } j=1 \text{ to } n \\
\quad & \text{if } (x[i] = y[j]) \quad c[i, j] \leftarrow c[i-1, j-1] + 1 \\
\quad & \text{else } \quad c[i, j] \leftarrow \max\{ c[i-1, j], c[i, j-1] \} \\
\end{align*}
\]

Time = \( \Theta(mn) \) = constant work per table entry. Space = \( \Theta(mn) \).

Reconstruction \( z = \text{LCS}(x, y) \)

**IDEA:** Compute the table bottom-up. Fill \( z \) backward.

**Observation:** \( c[i,j] \geq c[i-1,j] \) and \( c[i,j] \geq c[i,j-1] \)

**Proof Sketch:** We use a longer prefix, so there are more chars to be match.

LCS Reconstruction:
Set \( i=m; j=n; k=c[i,j] \)
While(\( k>0 \)){
if \( (c[i,j] \geq c[i-1,j] \) and \( c[i,j] \geq c[i,j-1] ) \) {
    \( z[k] = x[i] \); 
    \( i--; j--; k--; \)
}else if \( c[i,j] = c[i,j-1] \) {
    \( j--; \)
}else if \( c[i,j] = c[i-1,j] \) {
    \( i--; \)
}else{
    \( \}
}
Another idea – While filling \( c[i,j] \), add arrows to each cell \( c[i,j] \) specifying which neighboring cell \( c[i,j] \) it got its value.

- \( c[i,j].flag = "\" \) if \( c[i,j] = c[i-1,j-1]+1 \)
- \( c[i,j].flag = "↑" \) if \( c[i,j] = c[i-1,j] \)
- \( c[i,j].flag = "←" \) if \( c[i,j] = c[i,j-1] \)

Recursive algorithm for LCS
(just an example of what could go wrong)

\[
\text{LCS}(x, y, i, j) \quad \text{// we start the method with } i=|x|, \ j=|y|
\]
if ( \( i==0 \) or \( j=0 \) ) return 0
if \( x[i] = y[j] \)
then return \( \text{LCS}(x, y, i–1, j–1) + 1 \)
else return \( \max\{\text{LCS}(x, y, i–1, j), \ \text{LCS}(x, y, i, j–1)\} \)

Worst-case: \( x[i] \neq y[j] \), for all \( i,j \) in which case the algorithm evaluates two subproblems, each with only one parameter decremented.

Dynamic-programming hallmark #2

Overlapping subproblems

A recursive solution contains a “small” number of distinct subproblems repeated many times.

The number of distinct LCS subproblems for two strings of lengths \( m \) and \( n \) is only \( mn \).
Dynamic Programming:
Example 2: Rectilinear Steiner tree for points on 2 lines

- Input: a set of points (called “terminals”)
- A Steiner tree is a network that connects all of them, and of minimal length.
- This tree (in contrast to Minimum Spanning Tree) could have junctions (called Steiner Points) at point that are not terminals.
- The tree is rectilinear if edges are either parallel or orthogonal to each other.
- Special case: All terminals are on one of two parallel lines.

Solution - on whiteboard

Example 3: Edit distance

Given strings \( X, Y \), the edit distance \( ed(X, Y) \) between \( X \) and \( Y \) is defined as the minimum number of operations that we need to perform on \( X \), in order to obtain \( Y \).


Examples:
\[
\begin{align*}
ed("aaba", "aaba") &= 0 \\
ed("aaa", "abaa") &= 1 \\
ed("aaaa", "abaa") &= 1 \\
ed("baaa", ") &= 4 \\
ed("baaa", "aaab") &= 2 \\
\end{align*}
\]

Note that the term “distance” is a bit misleading: We need both the value (how many operations) as well as knowing which operations.

Example 3’:
``Priced” Edit distance \( ed(X, Y) \)

Assume also given
- \( InsCost \), - the cost of a single insertion into \( X \).
- \( DelCost \), - the cost of a single deletion from \( x \), and
- \( RepCost \), - the cost of replacing one character of \( x \) by a different character.

Definition: Given strings \( X, Y \), the edit distance \( ed(X, Y) \) between \( X \) and \( Y \) is the cheapest sequence of operations, starting on \( X \) and ending at \( Y \).

Problem: Compute \( ed(X, Y) \), (both the value and the optimal sequence of operations. )

Definition: \( c[i,j] = Cost( ed( X[1..i], Y[1..j] ) ) \).

Will first compute \( C[m,n] \). Then will recover the sequence.
Algorithm

**Memoization:** After computing a solution to a subproblem, store it in a table. Subsequent calls check the table to avoid redoing work.

\[
ed(X, Y) 
\]

for \( i = 0 \) to \( m \)  
\( c[i, 0] = i \text{ DelCost} \)

for \( j = 0 \) to \( n \)  
\( c[0, j] = j \text{ InsCost} \)

for \( i = 1 \) to \( m \)  
for \( j = 1 \) to \( n \)  
if \( X[i] = Y[j] \)  
then \( c[i, j] \leftarrow c[i-1, j-1] \)

else \( c[i, j] \leftarrow \min\{ \)
\( c[i-1, j] \) + \( \text{DelCost} \), \( c[i, j-1] \) + \( \text{InsCost} \), \( c[i-1, j-1] \) + \( \text{RepCost} \) \}
\}

Time = \( \Theta(mn) \) = constant work per table entry. Space = \( \Theta(mn) \).

Homework: Compute the sequence of operations.

Compute which characters in \( x \) matches which chars in \( y \).

**Polygonal Path - definition**

We define a polygonal path \( P = \{p_1 \ldots p_n\} \) where

- Each vertex \( p_i \) is a point in the plane,
- Vertex \( p_1 \) is the first vertex, \( p_n \) is the last,
- Vertex \( p_i \) is connected to the next vertex \( p_{i+1} \) by a straight segment.

\[ \]

**Good ways to measure distance between curves**

- Should not be effected by how curves are sampled
- Should reflect the “order” of the points along the curves.

\[ \]

\( P[1..i] \) is the polygonal line with the first \( i \) vertices of \( P \)

\( Q[1..j] \) is the polygonal line with the first \( j \) vertices of \( P \)
Definition of \( \text{Frechet}(P, Q, r) \)

Assume a person walks on \( P = \{p_1, \ldots, p_n\} \) while a dog walks on \( Q = \{q_1, \ldots, q_n\} \). \( r \) is the leash length (part of input).

The person starts at \( p_1 \) and ends at \( p_n \).

The dog starts at \( q_1 \) and ends at \( q_n \).

At each time stamp,
- either the person jumps to the next vertex
- or the dog jumps to the next vertex
- or both jumps to the next vertex
- Every instance they stop, we measure whether the distance between person ↔ dog (the length of the leash) \( \leq r \).

\( \text{Frechet}(P, Q, r) = \text{YES} \) if the answer is positive for all time stamps.

(if not, a longer leash is need. If yes, maybe a shorter one is sufficient.

So we could use binary search.

\[
\text{Frechet}(P, Q, r) \\
\text{//}\ c[1..n, 1..n] – boolean array \\
\text{//} c[i,j] = \text{Frechet}(P[1..i], Q[1..j], r) \\
\text{Init:} \\
c[1,1] = (|| p_1 - q_1 || \leq r) (YES/NO) \\
\text{For } i=2 \text{ to } n \\
c[i,1] = (|| p_i - q_1 || \leq r) \text{ AND } c[i-1,1] (YES/NO) \\
\text{For } j=2 \text{ to } n \\
c[1,j] = (|| p_1 - q_j || \leq r) \text{ AND } c[1,j-1] \\
\text{Return } c[n,n] \\
\]

Note – this is only the cost (that is the distance itself. We still need to find what is the series of steps that yield this cost

Another way to think about the \( \text{Frechet}(P, Q, r) \) problem

Create a graph \( G(V, E) \) where

\[ V = \{ v_{ij} \mid p_i, q_j \text{ are vertices of } P \text{ and } Q \} \]

That is, create a node for every pair of positions of (person,dog)

Edges connect two vertices that are legal move (person moves, dog moves, or both).

\[ v_{ij} \rightarrow v_{i+1,j} \quad v_{ij} \rightarrow v_{i,j+1} \quad v_{ij} \rightarrow v_{i+1,j+1} \]

Note \( |V|, |E| \) are both \( O(n^2) \).

Color each vertex are blue/red.

\( v_{ij} \) is blue iff \( |p_i - q_j| \leq r \).

\( v_{ij} \) is red iff \( |p_i - q_j| > r \).

Find a path from \( v_{1,1} \) to \( v_{n,n} \), that uses only blue vertices. (BSF)
Continuous Frechet($P, Q, \varepsilon$). Free Space Diagram.
(person, dog) could be anywhere along their paths, as long as distance between them $\leq \varepsilon$.

- The person and the dog move continuously along their trajectories. However, the distance between them must always be $\leq \varepsilon$ (given max leash distance).
- A “continuous” graph, which has vertices.
- Each point in the $(x, y)$ in the plane is associated with a pair person/dog location
- How? Person walks distance of $(starting at beginning of P)$
- Dog walks distance of $y$ from $(starting at beginning of Q)$
- For every segment $s_i$ of $P$, and every segment $e_j$ of $Q$ compute the (boundary of) all points representing acceptable locations (distance $\leq \varepsilon$). This takes $O(1)$ per a pair of segments.
- A point is black/white iff it represents a locations where distance acceptable. The resulting shape is called the Free-Space Diagram (term borrowed from robotics)
- Find a path inside the white region of the Free-Space diagram $O(n^2)$ time

Dynamic Time Warping $dtw(P,Q)$

Definition of $dtw(P,Q)$
Assume a person walks on $P=\{p_1, \ldots p_n\}$ while a dog walks on $Q=\{q_1, \ldots q_m\}$.

They person starts at $p_1$ and ends at $p_n$
They dog starts at $q_1$ and ends at $q_n$

At each time stamp,
• either the person jumps to the next vertex
• Or the dog jumps to the next vertex
• Or both jumps to the next vertex

• Every instance they stop, we measure the distance (the length of the leash) person$\leftrightarrow$dog.
• We sum the lengths of all leashes.
• $dtw(P,Q)$ is the smallest sum (over all possible sequences)

Motivation:

Definition of $dtw(P,Q)$
Assume a person walks on $P=\{p_1, \ldots p_n\}$ while a dog walks on $Q=\{q_1, \ldots q_m\}$.

Distance between trajectories enables finding nearest neighbor, and clustering

But two very similar trajectories might have vertices in very different places

Problem: Computing Dynamic Time Warping $dtw(P,Q)$ between polylines

Given 2 polygonal curves $P=\{p_1, \ldots p_n\}$ and $Q=\{q_1, \ldots q_m\}$.
The input is the locations of their vertices (e.g. GIS coordinates)

How similar are $P$ to $Q$?

Need to come up with a number $dtw(P,Q)$?
So if $dtw(P,Q) < dtw(P,Q')$, then $P$ is more similar to $Q$

DTW is used in
• Signal processing (speech reco)
• Signature verification
• Analysis of vehicles trajectories for roads networks
• Improving locations-based services
• Animals migrations patterns
• Stocks analysis
**Thm 1:**

Let \( c[i,j] = \text{dtw}( P[1..i], Q[1..j]) \).

Let \( || p_i - q_j || \) be the between the points \( p_i \) and \( q_j \)

That is, the length of the leash.

For every \( i>1 \), \( j>1 \)

\[ c[1,1] = || p_1 - q_1 || \]

\[ c[1,j] = c[1,j-1] + || p_1 - q_j || \]

\[ c[i,1] = c[i-1,1] + || p_i - q_1 || \]

**Thm 2:**

Assume at some time, the person is at \( p_i \) while dog at \( q_j \). Assume \( i>1 \) and \( j>1 \).

What (might have) happened one step ago?

Three possibilities

- Both person and the dog jumped (from \( p_{i-1} \) and from \( q_j \)) OR
- Person jumped from \( p_{i-1} \) to \( p_i \), dog stays at \( q_j \) OR
- Person stayed at \( p_i \), dog jumped from \( q_{j-1} \) to \( q_j \).

**Thm 2 cont:**

Let \( c[i,j] = \text{dtw}( P[1..i], Q[1..j]) \).

If \( i>1 \) and \( j>1 \) then

\[ c[i,j] = || p_i - q_j || + \]

\[ \min \{ \]

\[ c[i-1,j-1], // both jumps \]
\[ c[i-1,j], // person jumped from \( p_{i-1} \) to \( p_i \), dog stays at \( q_j \) \]
\[ c[i,j-1], // person stayed at \( p_i \), dog jumped from \( q_{j-1} \) to \( q_j \) \]

\} \]

Since we are not sure that when the person is at \( p_i \) the dog is at \( q_j \) we will compute all such pairs \( i,j \) – one of them must happened

**Algorithm for computing dtw(P,Q)**

Init according to Thm 1.

Fot \( i=2 \) to \( n \)

For \( j=2 \) to \( n \)

\[ c[i,j] = || p_i - q_j || + \]

\[ \min \{ \]

\[ c[i-1,j-1], // both jumps \]
\[ c[i-1,j], // person jumped from \( p_{i-1} \) to \( p_i \), dog stays at \( q_j \) \]
\[ c[i,j-1], // person stayed at \( p_i \), dog jumped from \( q_{j-1} \) to \( q_j \) \]

\} \]

Return \( c[n,n] \)

Note – this is only the cost (that is the distance itself. We still need to find what is the series of steps that yield this cost.
Ideas:
- Consider sub-polygons
- Assemble solutions to smaller problems to form solutions to larger polygons
- Let \( k \), the run-length, be an integer \( 1 \leq k \leq n \)
- A run of \( k \) vertices:
  - Start at \( p_i \), walks \( k \) vertices CCW (\( p_i, p_{i+1}, p_{i+2}, \ldots, p_{i+k-1} \))
  - Indexes are computed modulo \( n \), so a run of \( k = n \) starting at \( p_{n-1} \) is (\( p_{n-1}, p_0, p_1, p_2 \))
  - We should have written (\( p_i, p_{(i+1) \mod n}, p_{(i+2) \mod n} \)).
  - We omit the mod notation when there is no risk of confusion.
  - (\( p_0, \ldots, p_{n-1} \)) is the whole \( P \) (a run of length \( n \))

Finding Shortest Path in Large Graphs

- Phx
- I60
- Tucson
- Picacho Peak

Another Example of DP:
Polygon Triangulation

This is an interesting example because of the invariants - which cells are already calculated, and at which order to visit new cells so we only use already-computed values

- **Given** - a polygon \( P \) in the plane, it is given by specifying its vertices in the order they appear along the boundary (counter-clockwise).

- A chord is a segment fully inside \( P \), and connecting two vertices of \( P \).

- Triangulation of \( P \) is a partition of \( P \) into triangles using non-crossing chords. (add as many chords as possible, but don’t let them cross each other)

- **Goal**: Given \( P \), find a triangulation \( T \) that minimizes the cost

  \[
  \text{Cost} = \sum_{\Delta_i \in T} (\text{Area}(\Delta_i))^2
  \]

- (note that without the square, all triangulations are equally good)
- To present the algorithm, we could assume that \( P \) is convex (Otherwise we just ignore some chords.)
- **Fact**: The number of triangles is always \( n - 2 \), and number of chords is \( n - 3 \)

\[\text{Cost of triangulation} = \text{sum of squares of areas of its triangles}\]

Example c[4,10]:
- Consider sub-polygons
- Assemble solutions to smaller problems to form solutions to larger polygons
- Let \( k \), the run-length, be an integer \( 1 \leq k \leq n \)
- A run of \( k \) vertices:
  - Start at \( p_i \), walks \( k \) vertices CCW (\( p_i, p_{i+1}, \ldots, p_{i+k-1} \))

Define \( c[i,j] \) - the cost of opt triangulation of the sub polygon \( \{p_i, p_{i+1}, \ldots, p_j\} \). It assumes that the chord \( p_i, p_j \) is a part of the triangulation, so no chord starts at a vertex in \( \{p_i, p_{i+1}, \ldots, p_j\} \) connects to a vertex outside this sub-polygon.

- Note - when we compute \( c[i,j] \), we don’t know if this assumption is true. But it must be true for some pairs \( p_i, p_j \).

Define \( \Pi[i,j] \) is the third vertex of the triangle of the triangle \( \Delta \) that participates in this triangulation, and has \( p_i, p_j \) as an edge.

(in the example, \( \Pi[4,10] = 7 \), since \( p_7 \) is the third vertex of this triangle \( \Delta \).)
Example 6: Dynamic programming for TSP

- **Input**: A graph \( G(V,E) \), where \( w(v_i, v_j) \) is the cost of edge \( w(v_i, v_j) \).
- **Problem**: find a shortest path starting at vertex \( I \) and visits each node exactly once. Naive solution takes \( O(|V|^2) \).

  - **Properties**
    - if \( S = \{1, v_j\} \), then \( C(S, v_j) = w(1, v_j) \) (for \( k = 2, 3 \ldots n \))
    - for large sets \( S \) and any node \( v_k \in S \), lets analyze the path \( \pi \) of \( C(S, v_k) \). This is the shortest path that starts at 1, ends at \( v_j \), and visits all nodes of \( S \). Let \( v_i \) be the node of \( \pi \) just before node \( k \). Let \( \pi' \) denote the prefix of \( \pi \), from 1 till \( v_i \).
    - The cut-N-paste lemma implies that this \( \pi' \) is the shortest path that starts at 1, ends at \( v_j \), and visits all vertices of \( S - v_k \). So \( \pi' \) is the solution to \( C(S - v_k, v_k) \)

Problem: Need to find \( v_i \). So we check all options.

\[
C(S, v_j) = \min_{v_i \in S - v_k} \left\{ C(S - v_k, v_i) + w(v_i, v_j) \right\}
\]

### Algorithm

**Init (previous slide)**

- **For** \( k = 4 \ldots n-1 \) // length of run
  - **for** \( i = 0 \ldots n-1 \)
    - **j** = \( i + k - 1 \) // Note – not a loop
    - **for** \( t = i + 1 \) to j-1
      - \( A = c[i,t] + c[t,j] + \text{area}^2 \Delta(p_i p_j) \)
      - **if** \( A < c[i,j] \) then
        - \( c[i,j] = A \)
        - \( \Pi[i,j] = t \)

**Return** \( c[1,n] \)

**Homework**: Use \( \Pi \) to reconstruct the triangulation

https://www.geogebra.org/m/wq6snzgc

```
Remember
\( C[i,j] \) – the cost of opt triangulation of the sub polygon
  defined by the vertices \{ \( p_i, p_{i+1}, p_j \) \}
\( \Pi[i,j] \) contains the third vertex of the triangle that uses the edge \( p_i p_j \).

Init part of the algorithm:
// The init handles runs of length 0, 1 and 2.
For i=1 to n
For j=1 to n { c[i,j]=∞ ; \( \Pi[i,j]=\text{NULL} \} 
For i=1 to n { 
  \( c[i, i] = 0 ; \Pi[i, i] = p_i ; \)
  \( c[i, i+1] = 0 ; \Pi[i, i+1] = p_i ; \)
  \( c[i, i+2] = \text{Area}^2(\Delta p_i p_{i+1} p_{i+2}) \)
  \( \Pi[i, i+2] = p_{i+1} \)
// Every run of length 3 defines a triangle
}

https://www.geogebra.org/m/wq6snzgc
```
Does it worth the effort?

\[ O(2^n) \text{ vs } O(n!) \]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( 2^n )</th>
<th>( n! &gt; )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1024</td>
<td>3.6M</td>
</tr>
<tr>
<td>20</td>
<td>1M</td>
<td>1018</td>
</tr>
<tr>
<td>30</td>
<td>10^9</td>
<td>1035</td>
</tr>
</tbody>
</table>

Clustering Problem

\[
R + \text{Err}(\ell_1, \{p_1, p_2, \ldots, p_{i_1}\}) + R + \text{Err}(\ell_2, \{p_{i_1+1}, p_{i_1+2}, \ldots, p_{i_2}\}) + \ldots + R + \text{Err}(\ell_n, \{p_{i_n}, \ldots, p_p\})
\]

Given a point set \( P = \{p_1, \ldots, p_p\} \) sorted from left to right, and a cluster penalty \( R > 0 \).

Problem: Find a partition of \( P \) into \( k \) clusters such that the total clustering cost, \( tclt(P_1 \ldots P_k) \) is as small as possible. We define the total clustering cost, \( tclt(P_1 \ldots P_k) \) as the sum of \( k \) penalties \((k \cdot R)\), plus the sum of the fitting errors for the points in each cluster and the line the fit them best.

\[
tclt(P_1 \ldots P_k) = R + \text{Err}(\ell_1, \{p_1, p_2, \ldots, p_{i_1}\}) + R + \text{Err}(\ell_2, \{p_{i_1+1}, p_{i_1+2}, \ldots, p_{i_2}\}) + \ldots + R + \text{Err}(\ell_n, \{p_{i_n}, \ldots, p_p\})
\]

Note that if \( R = 0 \) (no penalty on new clusters) then the optimum clustering uses \( \frac{1}{2} \) runs:

\[
i_1, i_2, \ldots, i_n.
\]

In the example on top, \( k = 3 \), \( i_1 = 5 \), \( i_2 = 8 \)

Clustering a set of points

- A toy problem: Given an ordered set \( S = \{p_1, \ldots, p_p\} \) sorted by \( x \)-value where
- \( p_1 = (x_1, 0) \), cluster the set into \( n \) runs such that the errors of approximating each run is small, and the number of runs is small as well.
- A run \( C_k \) is a consecutive subset of points of \( S \). For example \( C_1 = \{p_1, p_2, \ldots, p_{i_1}\} \), \( C_2 = \{p_{i_1+1}, p_{i_1+2}, \ldots, p_{i_2}\} \), \( C_3 = \{p_{i_2+1}, p_{i_2+2}, \ldots, p_{i_3}\} \).
- The error is \( \frac{1}{2} \) (highest-lowest) in the cluster. Want to minimizes max error among all clusters.

\[
\max_{\text{cluster } C_k} \left\{ \frac{1}{2} \text{distortion}(p_{i_k}, y) - \min_{p_{i_{k+1}}} \right\}
\]

- Alternatively we could also minimizes the distance in each cluster to best-fit horizontal line.
- Find in each cluster the horizontal line that minimizes its max-distance to any point in the cluster.
- Problem: Find number of clusters, and find their boundaries.

Another application of DP: Clustering

(source: Kleinberg & Tardos 6.3)

DP meets in a clever way Divide & Conquer

Given a set of points \( S = \{p_1, \ldots, p_p\} \) and a line \( \ell = ax + b \) are given.

We could measure how well \( \ell \) fits \( S \) by checking, for each \( p_i \in S \), the vertical distance from \( \ell \), and sum the squares of these distances.

\[
\text{Err}(\ell, S) := \sum_{p_i \in S} (ax_i + b - y_i)^2 \text{ when } p_i = (x_i, y_i)
\]

Next, find the line \( y = a_{opt}x + b_{opt} \) that minimizes this error.

\[
0 = \frac{d}{db} \sum_{i} (y_i - ax_i - b)^2 = -2 \sum_{i} (y_i - ax_i - b) \\
0 = \frac{d}{dx} \sum_{i} (y_i - ax_i - b)^2 = -2 \sum_{i} (y_i - ax_i - b)
\]

\[
a_{opt} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} \quad \text{and} \quad b_{opt} = \frac{\sum y_i - a \sum x_i}{n}
\]

Def: Assume opt line \( \ell^* : y = a_{opt}x + b_{opt} \) is found.

Define: \( \text{Err}(S) := \sum_{i} (a_{opt}x_i + b_{opt} - y_i)^2 \) - fitting error.

Next, let \( S = \{p_1, p_2, \ldots, p_p\} \) be given, sorted from left to right. For every \( i, j \) \((i \leq j)\) we prepare an array \( \text{Defination } \text{Err}[i, j] \equiv \text{Err}(\{p_i, p_{i+1}, \ldots, p_j\}) \).

Claim (not critical): We can compute \( \text{Err}(\{p_i, p_{i+1}, \ldots, p_j\}) \) \( \forall i, j \) in time \( O(n^2) \).
In the example on top, points. Note that if total clustering cost, Problem: Find a partition of \( P \) into \( \frac{n}{2} \) clusters.

Given a point set \( P = \{p_1, p_2, \ldots, p_n\} \) sorted from left to right, and a cluster penalty \( R > 0 \).

Problem: Find a partition of \( P \) into \( \frac{n}{2} \) clusters.

\[
(p_1, p_2, \ldots, p_k), (p_{k+1}, p_{k+2}, \ldots, p_{k+\ell}), \ldots, (p_{n-\ell}, \ldots, p_n)
\]

and lines \( \ell_1, \ell_2, \ldots, \ell_k \) such that the total clustering cost, \( tct(p_1, p_n) \) is as small as possible. We define the total clustering cost, \( tct(p_1, p_n) \) as the sum of \( k \) penalties \( (k \cdot R) \), plus the sum of the fitting errors between the points in each cluster and the line the fit them best;

\[
tct(p_1, p_n) = \sum_{i=1}^{k} \ell_i\]

Note that if \( R = 0 \) (no penalty on new clusters) then the optimum clustering uses \( \frac{n}{2} \) runs:

\[
(p_1, p_2, \ldots, p_k), (p_{k+1}, p_{k+2}, \ldots, p_{k+\ell}), \ldots, (p_{n-\ell}, \ldots, p_n)
\]

If \( R \) is huge, then the optimum uses only one cluster, containing all the points.

In the example on top, \( k = 3, i_1 = 5, i_2 = 8 \)

Idea: The rightmost point of each cluster is the ‘cluster-head’, and this point has to pay the penalty \( R \).

---

### Summarizing

- The algorithm takes \( O(n^3) \) and \( O(n^2) \) space
- (for preprocessing \( d[j,i] \)
- Note – we did not discuss how to reconstruct the solution itself. We only calculated its cost

---

### Knapsack problem

Given:
1. Given a width \( W \) - this is the size of a (one dimensional) knapsack.
2. A set of segments \( S = \{s_1, \ldots, s_n\} \) on the x-axis,
3. The length of \( s_j \) is \( w_j \geq 0 \).
4. The revenue \( r_j \) of \( s_j \) is how much could we earn by fitting \( s_j \) into the “box”.

**Problem** find a subset \( S' \subseteq S \) so its total length \( \sum_{s_j \in S'} w_j \leq W \), and the total revenue \( \sum_{s_j \in S'} r_j \) is maximized.

Could fit \( g, k, i \), total length \( = 2 + 4 + 3 = 9 \)

Revenue = \( 12$ + 10$ + 11$ = 33$
Knapsack problem

- The problem is NP-hard. But easier to solve if we assume all length $w_i$ are integers.
- Let $S_i$ be the set of segments $S_i = \{s_1…s_j\}$. $S_0 = \emptyset$.
- $c(i, j)$ is the opt revenue solution to the knapsack problem when segment is only picked from $S_i$, and the total length is at most $\sum w_i$ is at most $j$ (here $j \leq W$ is an int).
- The inner loop checks $i=0,1,2..n$. That is, by increasing $i$, we could use more segments.
- In other words, we only could use segments from $S_i$ (instead of $S$) and we need to fit them into a narrow knapsack of width $j$ (instead of $W$).
- Fix knapsack width $j$. Assume $c(i, j)$ is already calculated for all $1 \leq i \leq n$.
- The outer loop of the algorithm checks wider and wider knapsacks.
- The inner loop checks $i=0,1,2..n$. That is, by increasing $i$, we could use more segments.
- Each segment is used at most once.

- Observation: but should we use $s_i$ (for $[1..j]$)?
- If we do use $s_i$, then $c(i, j) = c(i - 1, j)$
- If we don’t, then $c(i, j) = c(i - 1, j - w_i)$
- For $i = 0$ clearly $c(0, j) = 0$ (since $S_0 = \emptyset$, we cannot use any segment)
- Increasing $i$ from $i - 1$ to $i$ implies that for the knapsack $[0..j]$ we could (allowed to) use $s_i$ - but should we use $s_i$ (for $[1..j]$)?

Algorithm
1. Init $c(0, j) = 0$. $\forall j$
2. For $j = 1 \ldots W$
   2.1. For $i = 1 \ldots n$
      2.1.1. $c(i, j) = \max\{c(i - 1, j - 1), \ r_i + c(i - 1, j - w_i)\}$
      2.1.2. %Of course, keep $\Pi[1..n, 1..W]$ to construct the opt solution.

- Let $S_i$ be the set of segments $S_i = \{s_1…s_j\}$. $S_0 = \emptyset$.
- Notation: $[0..j]$ is a knapsack of length $j$. The original knapsack is $[0..W]$.
- $c(i, j)$ is the opt revenue for the knapsack problem, with two restriction:
  1. Segment are picked on from from $S_i$.
  2. The total length is at most $\sum w_i$ is at most $j$ (here $j \leq W$ is an int). That is, fit into the knapsack $[0..j]$.
- Fix knapsack width $j$. Assume $c(i, j)$ is already calculated for all $1 \leq j' \leq j$ (narrower knapsacks).
- For $i = 0$ clearly $c(0, j) = 0$ (since $S_0 = \emptyset$, we cannot use any segment)
- Increasing $i$ from $i - 1$ to $i$ implies that for the knapsack $[0..j]$ we could (allowed to) use $s_i$ - but should we use $s_i$ (for $[1..j]$)?

- If we do use $s_i$, at this point $(j,0)$, then we place $s_i$ so its right endpoint at $(j,0)$. So its left endpoint is at $(j - w_i, 0)$, and the max revenue is $r_i + c(i - 1, j - w_i)$

- Observation: $c(i, j) = \max\{c(i - 1, j), \ r_i + c(i - 1, j - w_i)\}$.