Chapter 6
Dynamic Programming

Algorithmic Paradigms

Greedy. Build up a solution incrementally, myopically optimizing some local criterion.

Divide-and-conquer. Break up a problem into sub-problems, solve each sub-problem independently, and combine solution to sub-problems to form solution to original problem.

Dynamic programming. Break up a problem into a series of overlapping sub-problems, and build up solutions to larger and larger sub-problems.

Dynamic Programming History

Bellman. [1950s] Pioneered the systematic study of dynamic programming.

Etymology.
- Dynamic programming = planning over time.
- Secretary of Defense was hostile to mathematical research.
- Bellman sought an impressive name to avoid confrontation.

"it's impossible to use dynamic in a pejorative sense"
"something not even a Congressman could object to"


Dynamic Programming Applications

Areas.
- Bioinformatics.
- Control theory.
- Information theory.
- Operations research.
- Computer science: theory, graphics, AI, compilers, systems, ...

Some famous dynamic programming algorithms.
- Unix diff for comparing two files.
- Viterbi for hidden Markov models.
- Smith-Waterman for genetic sequence alignment.
- Bellman-Ford for shortest path routing in networks.
- Cocke-Kasami-Younger for parsing context free grammars.
6.1 Weighted Interval Scheduling

Weighted interval scheduling problem.
- Job $j$ starts at $s_j$, finishes at $f_j$, and has weight or value $v_j$.
- Two jobs compatible if they don’t overlap.
- Goal: find maximum weight subset of mutually compatible jobs.

Unweighted Interval Scheduling Review

Recall. Greedy algorithm works if all weights are 1.
- Consider jobs in ascending order of finish time.
- Add job to subset if it is compatible with previously chosen jobs.

Observation. Greedy algorithm can fail spectacularly if arbitrary weights are allowed.

Weighted Interval Scheduling

Notation. Label jobs by finishing time: $f_1 \leq f_2 \leq \ldots \leq f_n$.
Def. $p(j)$ = largest index $i < j$ such that job $i$ is compatible with $j$.

Ex: $p(8) = 5$, $p(7) = 3$, $p(2) = 0$. 

Weighted interval scheduling problem. 
- Job $j$ starts at $s_j$, finishes at $f_j$, and has weight or value $v_j$.
- Two jobs compatible if they don’t overlap.
- Goal: find maximum weight subset of mutually compatible jobs.
Dynamic Programming: Binary Choice

Notation. \( OPT(j) \) = value of optimal solution to the problem consisting of job requests 1, 2, ..., j.

- Case 1: \( \text{OPT} \) selects job j.
  - collect profit \( v_j \)
  - can’t use incompatible jobs \{ p(j) + 1, p(j) + 2, ..., j - 1 \}
  - must include optimal solution to problem consisting of remaining compatible jobs 1, 2, ..., \( p(j) \)

- Case 2: \( \text{OPT} \) does not select job j.
  - must include optimal solution to problem consisting of remaining compatible jobs 1, 2, ..., j-1

\[
OPT(j) = \begin{cases} 
0 & \text{if } j = 0 \\
\max \{ v_j + OPT(p(j)), OPT(j-1) \} & \text{otherwise} 
\end{cases}
\]

Weighted Interval Scheduling: Memoization

Memoization. Store results of each sub-problem in a cache; lookup as needed.

\[
\begin{align*}
\text{Input: } & n, s_1, \ldots, s_n, f_1, \ldots, f_n, v_1, \ldots, v_n \\
\text{Sort jobs by finish times so that } & f_1 \leq f_2 \leq \ldots \leq f_n. \\
\text{Compute } & p(1), p(2), \ldots, p(n) \\
\text{Compute-Memoize}(j) \{ \\
\quad & \text{if } (j = 0) \text{ return } 0 \\
\quad & \text{else return } \max(v_j + \text{Compute-Memoize}(p(j)), \text{Compute-Memoize}(j-1)) \\
\}
\end{align*}
\]
Weighted Interval Scheduling: Running Time

Claim. Memoized version of algorithm takes $O(n \log n)$ time.
- Sort by finish time: $O(n \log n)$.
- Computing $p()$: $O(n \log n)$ via sorting by start time.

- M-Compute-Opt(j): each invocation takes $O(1)$ time and either
  - (i) returns an existing value $M[j]$
  - (ii) fills in one new entry $M[j]$ and makes two recursive calls

- Progress measure $\Phi = \# \text{ nonempty entries of } M[]$.
  - initially $\Phi = 0$, throughout $\Phi \leq n$.
  - (ii) increases $\Phi$ by 1 $\Rightarrow$ at most $2n$ recursive calls.

- Overall running time of M-Compute-Opt(n) is $O(n)$. •

Remark. $O(n)$ if jobs are pre-sorted by start and finish times.

Weighted Interval Scheduling: Finding a Solution

Q. Dynamic programming algorithms computes optimal value. What if we want the solution itself?
A. Do some post-processing.

Run M-Compute-Opt(n)
Run Find-Solution(n)

Find-Solution(j) {
  if (j = 0)
    output nothing
  else if ($v_j + M[p(j)] > M[j-1]$)
    print j
    Find-Solution(p(j))
  else
    Find-Solution(j-1)
}

- # of recursive calls $\leq n \Rightarrow O(n)$.

Weighted Interval Scheduling: Bottom-Up

Bottom-up dynamic programming. Unwind recursion.

Input: n, $s_1, ..., s_n$, $f_1, ..., f_n$, $v_1, ..., v_n$
Sort jobs by finish times so that $f_1 \leq f_2 \leq ... \leq f_n$.
Compute $p(1), p(2), ..., p(n)$
Iterative-Compute-Opt {
  $M[0] = 0$
  for j = 1 to n
    $M[j] = \max(v_j + M[p(j)], M[j-1])$
}

6.3 Segmented Least Squares
Segmented Least Squares

Least squares.
- Foundational problem in statistic and numerical analysis.
- Given $n$ points in the plane: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Find a line $y = ax + b$ that minimizes the sum of the squared error:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Solution. Calculus $\Rightarrow$ min error is achieved when

$$a = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}, \quad b = \frac{\sum y_i - a \sum x_i}{n}$$

Dynamic Programming: Multiway Choice

Notation.
- $\text{OPT}(j) =$ minimum cost for points $p_1, p_{i+1}, \ldots, p_j$.
- $e(i, j) =$ minimum sum of squares for points $p_i, p_{i+1}, \ldots, p_j$.

To compute $\text{OPT}(j)$:
- Last segment uses points $p_i, p_{i+1}, \ldots, p_j$ for some $i$.
- Cost $= e(i, j) + c + \text{OPT}(i-1)$.

$$\text{OPT}(j) = \begin{cases} 0 & \text{if } j = 0 \\ \min_{1 \leq i \leq j} \{ e(i, j) + c + \text{OPT}(i-1) \} & \text{otherwise} \end{cases}$$

Segmented least squares.
- Points lie roughly on a sequence of several line segments.
- Given $n$ points in the plane $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ with $x_1 < x_2 < \ldots < x_n$, find a sequence of lines that minimizes $f(x)$.

Q. What's a reasonable choice for $f(x)$ to balance accuracy and parsimony?

$\uparrow$ goodness of fit
\numberoflines

Q. What's a reasonable choice for $f(x)$ to balance accuracy and parsimony?

$\uparrow$ goodness of fit
\numberoflines
Segmented Least Squares: Algorithm

**INPUT:** $n$, $P_1, \ldots, P_n$, $c$

```c
Segmented-Least-Squares() {
    M[0] = 0
    for $j = 1$ to $n$
        for $i = 1$ to $j$
            compute the least square error $e_{ij}$ for the segment $P_i, \ldots, P_j$
        for $j = 1$ to $n$
            $M[j] = \min_{1 \leq i \leq j} (e_{ij} + c + M[i-1])$
    return $M[n]$
}
```

Running time. $O(n^3)$. can be improved to $O(n^2)$ by pre-computing various statistics.

- Bottleneck = computing $e(i, j)$ for $O(n^2)$ pairs, $O(n)$ per pair using previous formula.

6.4 Knapsack Problem

**Knapsack problem.**
- Given $n$ objects and a "knapsack."
- Item $i$ weighs $w_i > 0$ kilograms and has value $v_i > 0$.
- Knapsack has capacity of $W$ kilograms.
- Goal: fill knapsack so as to maximize total value.

**Ex:** $\{ 3, 4 \}$ has value 40.

<table>
<thead>
<tr>
<th>#</th>
<th>value</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>7</td>
</tr>
</tbody>
</table>

**Greedy:** repeatedly add item with maximum ratio $v_i / w_i$.
**Ex:** $\{ 5, 2, 1 \}$ achieves only value $= 35 \Rightarrow$ greedy not optimal.

Dynamic Programming: False Start

**Def.** $OPT(i) = \text{max profit subset of items } 1, \ldots, i$.

- **Case 1:** $OPT$ does not select item $i$.
  - $OPT$ selects best of $\{ 1, 2, \ldots, i-1 \}$
- **Case 2:** $OPT$ selects item $i$.
  - accepting item $i$ does not immediately imply that we will have to reject other items
  - without knowing what other items were selected before $i,$ we don’t even know if we have enough room for $i$

**Conclusion.** Need more sub-problems!
Dynamic Programming: Adding a New Variable

Def. $OPT(i, w) =$ max profit subset of items 1, ..., i with weight limit w.

- Case 1: OPT does not select item i.
  - OPT selects best of { 1, 2, ..., i-1 } using weight limit w

- Case 2: OPT selects item i.
  - new weight limit = $w_i$
  - OPT selects best of { 1, 2, ..., i-1 } using this new weight limit

$$OPT(i, w) = \begin{cases} 
0 & \text{if } i = 0 \\
OPT(i-1, w) & \text{if } w_i > w \\
\max \{ OPT(i-1, w), \ v_i + OPT(i-1, w-w_i) \} & \text{otherwise}
\end{cases}$$

Knapsack Problem: Bottom-Up

Knapsack. Fill up an n-by-W array.

Input: $n$, $W$, $w_1$, ..., $w_n$, $v_1$, ..., $v_n$

for $w = 0$ to $W$
  $M[0, w] = 0$

for $i = 1$ to $n$
  for $w = 1$ to $W$
    if $(w_i > w)$
      $M[i, w] = M[i-1, w]$
    else
      $M[i, w] = \max \{ M[i-1, w], v_i + M[i-1, w-w_i] \}$
  return $M[n, W]$

Knapsack Problem: Running Time

Running time. $\Theta(nW)$.

- Not polynomial in input size!
  - "Pseudo-polynomial."
- Decision version of Knapsack is NP-complete. [Chapter 8]

Knapsack approximation algorithm. There exists a poly-time algorithm that produces a feasible solution that has value within 0.01% of optimum. [Section 11.8]
RNA Secondary Structure

**RNA.** String $B = b_1b_2\ldots b_n$ over alphabet \{ A, C, G, U \}.

**Secondary structure.** RNA is single-stranded so it tends to loop back and form base pairs with itself. This structure is essential for understanding behavior of molecule.

Ex: GUCGAUGAGCGAAUGUAACAACGUGGCUACGGCGAGA

complementary base pairs: A-U, C-G

**Free energy.** Usual hypothesis is that an RNA molecule will form the secondary structure with the optimum total free energy.

**Goal.** Given an RNA molecule $B = b_1b_2\ldots b_n$, find a secondary structure $S$ that maximizes the number of base pairs.

**Examples.**

RNA Secondary Structure: Examples

- [Watson-Crick.] $S$ is a matching and each pair in $S$ is a Watson-Crick complement: A-U, U-A, C-G, or G-C.
- [No sharp turns.] The ends of each pair are separated by at least 4 intervening bases. If $(b_i, b_j) \in S$, then $i < j - 4$.
- [Non-crossing.] If $(b_i, b_j)$ and $(b_k, b_l)$ are two pairs in $S$, then we cannot have $i < k < j < l$.

- ok
- sharp turn
- crossing
RNA Secondary Structure: Subproblems

First attempt. \( \text{OPT}(j) = \) maximum number of base pairs in a secondary structure of the substring \( b_1b_2...b_j \).

Difficulty. Results in two sub-problems.
- Finding secondary structure in: \( b_1b_2...b_{t-1} \).
- Finding secondary structure in: \( b_{t+1}b_{t+2}...b_{n-1} \).

Dynamic Programming Over Intervals

Notation. \( \text{OPT}(i, j) = \) maximum number of base pairs in a secondary structure of the substring \( b_ib_{i+1}...b_j \).

- Case 1. If \( i \geq j - 4 \).
  - \( \text{OPT}(i, j) = 0 \) by no-sharp turns condition.
- Case 2. Base \( b_j \) is not involved in a pair.
  - \( \text{OPT}(i, j) = \text{OPT}(i, j-1) \)
- Case 3. Base \( b_j \) pairs with \( b_t \) for some \( i \leq t < j - 4 \).
  - non-crossing constraint decouples resulting sub-problems
  - \( \text{OPT}(i, j) = 1 + \max \{ \text{OPT}(i, t-1) + \text{OPT}(t+1, j-1) \} \)

Remark. Same core idea in CKY algorithm to parse context-free grammars.

Bottom Up Dynamic Programming Over Intervals

Q. What order to solve the sub-problems?
A. Do shortest intervals first.

Running time. \( O(n^3) \).

Dynamic Programming Summary

Recipe.
- Characterize structure of problem.
- Recursively define value of optimal solution.
- Compute value of optimal solution.
- Construct optimal solution from computed information.

Dynamic programming techniques.
- Binary choice: weighted interval scheduling.
- Multi-way choice: segmented least squares.
- Adding a new variable: knapsack.
- Dynamic programming over intervals: RNA secondary structure.

Top-down vs. bottom-up: different people have different intuitions.
6.6 Sequence Alignment

String Similarity

How similar are two strings?
- occurrance
- occurrence

Applications.
- Basis for Unix diff.
- Speech recognition.
- Computational biology.

- Gap penalty $\delta$; mismatch penalty $\alpha_{pq}$.
- Cost = sum of gap and mismatch penalties.

Sequence Alignment

Goal: Given two strings $X = x_1 x_2 \ldots x_m$ and $Y = y_1 y_2 \ldots y_n$ find alignment of minimum cost.

Def. An alignment $M$ is a set of ordered pairs $x_i-y_j$ such that each item occurs in at most one pair and no crossings.

Def. The pair $x_i-y_j$ and $x_{i'}-y_{j'}$ cross if $i < i'$, but $j > j'$.

Ex: CTACCG vs. TACATG.
Sol: $M = x_2-y_1, x_3-y_2, x_4-y_3, x_5-y_4, x_6-y_6$. 

\[
\text{cost}(M) = \sum_{(i,j) \in M} \alpha_{x_i y_j} + \sum_{j \in M \text{ unmatched}} \delta + \sum_{j \in y \text{ unmatched}} \delta
\]
Sequence Alignment: Problem Structure

**Def.** $OPT(i, j) = \text{min cost of aligning strings } x_1 x_2 \ldots x_i \text{ and } y_1 y_2 \ldots y_j$.

- **Case 1:** $OPT$ matches $x_i$-$y_j$.
  - Pay mismatch for $x_i$-$y_j$ + min cost of aligning two strings $x_1 x_2 \ldots x_{i-1}$ and $y_1 y_2 \ldots y_{j-1}$.
- **Case 2a:** $OPT$ leaves $x_i$ unmatched.
  - Pay gap for $x_i$ and min cost of aligning $x_1 x_2 \ldots x_{i-1}$ and $y_1 y_2 \ldots y_j$.
- **Case 2b:** $OPT$ leaves $y_j$ unmatched.
  - Pay gap for $y_j$ and min cost of aligning $x_1 x_2 \ldots x_i$ and $y_1 y_2 \ldots y_{j-1}$.

$$OPT(i, j) = \begin{cases} 
  j\delta & \text{if } i = 0 \\
  \alpha_{x_i y_j} + OPT(i-1, j-1) & \text{otherwise} \\
  \delta + OPT(i-1, j) & \text{if } j = 0 \\
  \delta + OPT(i, j-1) & \text{otherwise} 
\end{cases}$$

Sequence Alignment: Algorithm

```java
Sequence-Alignment(m, n, x_1 x_2 \ldots x_m, y_1 y_2 \ldots y_n, \delta, \alpha) {
    for i = 0 to m
        M[0, i] = i \delta
    for j = 0 to n
        M[j, 0] = j \delta
    for i = 1 to m
        for j = 1 to n
            M[i, j] = min(\alpha[x_i, y_j] + M[i-1, j-1], \delta + M[i-1, j], \delta + M[i, j-1])
    return M[m, n]
}
```

Analysis. $\Theta(mn)$ time and space.

English words or sentences: $m, n \leq 10$.

Computational biology: $m = n = 100,000$. 10 billions ops OK, but 10GB array?

Sequence Alignment: Linear Space

**Q.** Can we avoid using quadratic space?

**Easy.** Optimal value in $O(m + n)$ space and $O(mn)$ time.

- Compute $OPT(i, \cdot)$ from $OPT(i-1, \cdot)$.
- No longer a simple way to recover alignment itself.

**Theorem.** [Hirschberg 1975] Optimal alignment in $O(m + n)$ space and $O(mn)$ time.

- Clever combination of divide-and-conquer and dynamic programming.
- Inspired by idea of Savitch from complexity theory.

6.7 Sequence Alignment in Linear Space
**Edit distance graph.**
- Let $f(i, j)$ be shortest path from (0,0) to (i, j).
- Observation: $f(i, j) = \text{OPT}(i, j)$.

**Sequence Alignment: Linear Space**

- Can compute $f(\cdot, j)$ for any $j$ in $O(mn)$ time and $O(m + n)$ space.

---

**Edit distance graph.**
- Let $g(i, j)$ be shortest path from (i, j) to (m, n).
- Can compute $g(\cdot, j)$ for any $j$ in $O(mn)$ time and $O(m + n)$ space.
Observation 1. The cost of the shortest path that uses \((i, j)\) is \(f(i, j) + g(i, j)\).

Sequence Alignment: Linear Space

```
i-j
m-n
x1 x2
y1 y2 y3 y4 y5 y6
```

Observation 2. Let \(q\) be an index that minimizes \(f(q, n/2) + g(q, n/2)\). Then, the shortest path from \((0, 0)\) to \((m, n)\) uses \((q, n/2)\).

```
i-j
m-n
x1 x2
y1 y2 y3 y4 y5 y6
```

Divide: find index \(q\) that minimizes \(f(q, n/2) + g(q, n/2)\) using DP.
- Align \(x_q\) and \(y_{n/2}\).
Conquer: recursively compute optimal alignment in each piece.

Theorem. Let \(T(m, n) = \max\) running time of algorithm on strings of length at most \(m\) and \(n\). \(T(m, n) = O(mn \log n)\).

Remark. Analysis is not tight because two sub-problems are of size \((q, n/2)\) and \((m - q, n/2)\). In next slide, we save \(\log n\) factor.
Theorem. Let $T(m, n) = \max$ running time of algorithm on strings of length $m$ and $n$. $T(m, n) = O(mn)$.

Pf. (by induction on $n$)
- $O(mn)$ time to compute $f(\cdot, n/2)$ and $g(\cdot, n/2)$ and find index $q$.
- $T(q, n/2) + T(m - q, n/2)$ time for two recursive calls.
- Choose constant $c$ so that:

\[
\begin{align*}
T(m, 2) &\leq cm \\
T(2, n) &\leq cn \\
T(m, n) &\leq cmn + T(q, n/2) + T(m - q, n/2)
\end{align*}
\]

- Base cases: $m = 2$ or $n = 2$.
- Inductive hypothesis: $T(m, n) \leq 2cmn$. 

\[
\begin{align*}
T(m, n) &\leq T(q, n/2) + T(m - q, n/2) + cmn \\
&\leq 2cqn/2 + 2c(m - q)n/2 + cmn \\
&= cqn + cmn - cqn + cmn \\
&= 2cmn
\end{align*}
\]