

Longest Common Subsequence Problem

- **Definition** (Common Subsequence)

Let $V = v_1v_2 \dots v_n$ and $W = w_1w_2 \dots w_m$. A common subsequence of V and W , U , is a sequence which is a subsequence of V and respectively W .

- For example, if $V = \text{ATCTGAT}$ and $W = \text{TGCATA}$, then **TCTA** is a common subsequence of V and W .

- **Definition** ($s(V, W)$)

The length of a longest common subsequence (LCS) of V and W is denoted by $s(V, W)$.

Edit Distance

- Definition (Edit Distance)
- The minimum number of insertions and deletions needed to transform from a sequence V into another sequence W is called the edit distance from V into W , denoted by $d(V, W)$.
- For example, if $V = \text{ATCTGAT}$ and $W = \text{TGCATA}$, then $d(V, W) = 5$. (?)
It takes three deletions from V and insert two letters to obtain W .
- **Proposition** For two sequences V and W ,
$$d(V, W) = |V| + |W| - s(V, W).$$
- **Note.** To find edit distance, it suffices to find LCS of V and W .

Dynamic Algorithm of LCS

- Let $V = v_1v_2 \dots v_n$, $W = w_1w_2 \dots w_m$, $V_i = v_1v_2 \dots v_i$ and $W_j = w_1w_2 \dots w_j$.
- Let $s(V_i, W_j) = s_{i,j}$ where $1 \leq i \leq n$ and $1 \leq j \leq m$.

Algorithm

- (1) $s_{i,j}$ is equal to $s_{i,j-1}$ and $s_{i-1,j}$; and
- (2) $s_{i,j} = s_{i-1,j-1} + 1$ if $v_i = w_j$. Initial condition: $s_{i,0} = s_{0,j} = 0$.
- We can use an $(n+1) \times (m+1)$ array to depict the algorithm, see an example where $V = \text{TGCATA}$ and $W = \text{ATCTGAT}$.
- It takes $O(nm)$ time and space.

Combinatorial Relation

- Consider a permutation π of $\{1, 2, \dots, n\}$. π is also known as a permutation of order n .

Problem Find a longest increasing subsequence of π .

- The problem can be solved by finding the LCS of π and the identity permutation of order n , that is, the permutation with cycle structure $(1)(2) \dots (n)$.
- The representation of $\pi = \begin{bmatrix} 1 & 2 & 3 & \dots & n \\ x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix}$

Review of DM

- Theorem

Every sequence of real numbers with $n^2 + 1$ terms contains a monotone subsequence with at least $n + 1$ terms.

Proof. By using *pigeon hole principle*.

- Pigeon hole principle is also known as “average value” principle. It is very useful in showing the **existence** of certain configurations.



A Better Algorithm (*Using Young Tableaux*)

- Definition

A partition of a positive integer n is a sequence of positive integers $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_l$ such that their sum is equal to n . For convenience, we use

$\lambda = (\lambda_1, \lambda_2, \dots, \lambda_l)$ to denote the partition of n , written as $\lambda \vdash n$.

For example, $\lambda = (4, 2, 2, 1)$ is a partition of $n = 9$.



Young Diagram

- Definition (Young Diagram)

The Young diagram of shape $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_l)$ for a positive integer n is an array of n cells arranged into l left-justified rows with row i containing λ_i cells for $i = 1, 2, \dots, l$.

For example, the following “**shape**” is a Young diagram of $(4, 2, 1, 1)$.

1 2 5 8

4 7

6

9

Young Tableaux

- Definition (Young Tableau)

A Young tableau is array obtained by replacing the cells of the Young diagram with the numbers 1, 2, ..., n . A tableau is standard if its rows and columns are increasing sequences. The following table is an example for $n = 9$.

1	2	3	8
4	5		
6	7		
9			

Permutation and (P, Q) - Tableaux

(*) Every permutation can be represented by a pair of tableaux, and there is a bijection between permutations and pairs of tableaux.

- The above observation was first proposed by Robinson in 1938, and then Schensted in 1961. Finally, in 1970, Knuth generalized the algorithm for the case of LCS problem.
- It is known today as the RSK algorithm.
- We focus on finding the above mentioned algorithm here in this lecture.

Example First

- $\pi = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 7 & 2 & 8 & 1 & 3 & 4 & 10 & 6 & 9 & 5 \end{bmatrix}$
= (7 2 8 1 3 4 10 6 9 5)

- By using the algorithm, we shall obtain a pair of tableaux: P in the left and Q to the right. (Find them yourself.)

1	3	4	5	9
2	6	10		
7	8			

1	3	6	7	9
2	5	8		
4	10			

RSK Algorithm

- A partial tableau P is a Young diagram with distinct entries where rows and columns increase. (The following steps find P .)
- For a row R and an element x , define x_R^+ as the smallest element of R greater than x , and x_R^- as the largest element of R smaller than x . For x not in P , let the row insertion of x into P be obtained by the following steps.

$R \leftarrow$ the first row of P

while x is less than some element of the row R

Replace x_R^+ by x in R

$R \leftarrow$ next row (down)

Place x at the end of row R if x is greater than every element of R

Finding (P, Q)

- Keep the record of inserting new cells will give us the tableau Q .
- In fact, this algorithm provides a sequence of pairs: $(P_0, Q_0) = (\Phi, \Phi)$, (P_1, Q_1) , (P_2, Q_2) , ..., $(P_n, Q_n) = (P, Q)$.
- (P_k, Q_k) is obtained from inserting x_k into P_k and put k in the corresponding cell to make sure they have the same “size”.
- Here is another example. $\Pi = (4 \ 2 \ 3 \ 6 \ 5 \ 1 \ 7)$, (P, Q) are:

1 3 5 7

2 6

4

1 3 4 7

2 5

6

Find Permutation from (P, Q)

- The permutation will be obtained starting from the last element x_n , and the $x_{n-1}, x_{n-2}, \dots, x_1$.
- Consider the example in previous page.
 - (1) Since 7 is in the 7th position of Q_7 , and it is in the first row, we conclude that $x_7 = 7$.
 - (2) Delete 7 in P and its corresponding position in Q, we have (P_6, Q_6) . Now, the 6th position in Q_6 is at the bottom and in P, it is filled with 4. By looking up the column, we conclude that $x_6 = 1$.
 - (3) Delete 1 in P_6 (move 2 and 4 up), we consider (P_5, Q_5) . Now the 5th position of Q is filled with 6. In this case 5 is x_R^- , then $x_5 = 5$. **The process continues until the permutation is decided.**

Main Result

- Theorem

The map $\pi \rightarrow (P, Q)$ is a bijection between elements of S_n and pairs of standard tableaux of the shape $\lambda \vdash n$.

Proof. Given a permutation π , the pair of standard tableaux can be obtained by using the RSK algorithm mentioned in previous pages. Therefore, it suffices to prove that a pair of tableaux (P, Q) from $\lambda \vdash n$ can determine a unique permutation. The idea of finding the permutation $(x_1 x_2 \dots x_n)$ is by recursive argument starting from determining x_n , and then $x_{n-1}, x_{n-2}, \dots, x_1$. Mainly, assume that (P_k, Q_k) has been constructed for $k = n, n-1, \dots, 2$, we will find (P_{k-1}, Q_{k-1}) and x_k .

About the length of LCS

- *Exercise 2*

- (1) Give a detail proof of the **theorem** mentioned in last page.
- (2) Let P be the standard tableau obtained from π in the above theorem. **Prove that** the length of the longest increasing subsequence of π is the length of the first row of P .

Note. You may explain your idea by using examples.

Bonus: Given a permutation (randomly) of S_n , find the expectation length of an increasing subsequence.