Longest Common Subsequence Problem (LCS)

Definition (Common subsequence)

Let $V = \langle v_1, v_2, \cdots, v_m \rangle$ and $W = \langle w_1, w_2, \cdots, w_n \rangle$. A common subsequence of V and W is a sequence of indices $1 \le i_1 \le i_2 \le \cdots \le i_k \le m$ chosen from V and a sequence of indices $1 \le j_1 \le j_2 \le \cdots \le j_k \le n$ chosen from W such that $v_{i_t} = w_{j_t}$ for $t = 1, 2, \cdots, k$.

e.g. $V = A\,T\,C\,T\,G\,A\,T = \langle v_1, v_2, \cdots, v_7 \rangle$

$W = T\,G\,C\,A\,T\,A = \langle w_1, w_2, \cdots, w_6 \rangle$

$i_1 = 2,\ i_2 = 3,\ i_3 = 4,\ i_4 = 6 \Rightarrow \langle v_2, v_3, v_4, v_6 \rangle$

$j_1 = 1,\ j_2 = 3,\ j_3 = 5,\ j_4 = 6 \Rightarrow \langle w_1, w_3, w_5, w_6 \rangle$

$v_{i_1} = v_2 = w_1 = w_{j_1}$, $v_{i_2} = v_3 = w_3 = w_{j_2}$ ...

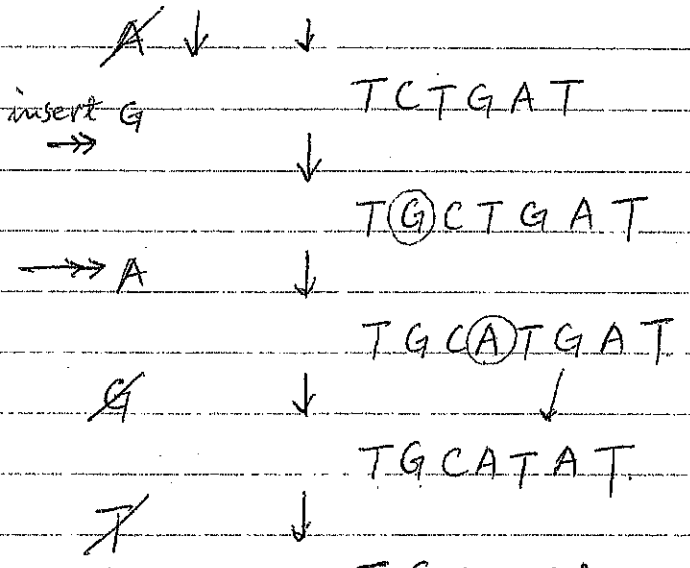
Definition ($s(V, W)$)

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

Definition ($d(V,W)$)

The minimum number of insertions and deletions needed to transform from V into W is called the edit distance from V into W, denoted by $d(V,W)$.

e.g. $V = ATCTGAT$

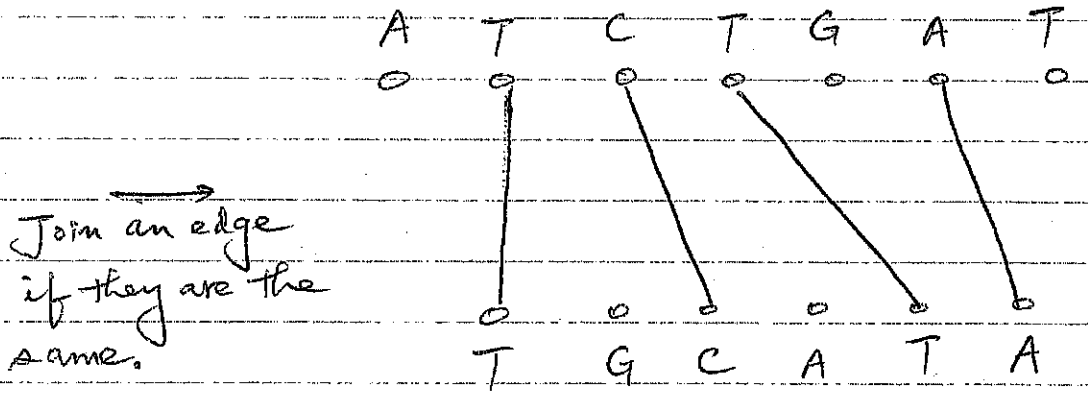


It takes two insertions and three deletions to transform V into W. Hence, $d(V,W) \leq 5$. In fact, we can verify the following property.

Proposition (Fact 1)

$d(V,W) = |V| + |W| - 2s(V,W)$, $|V|$ and $|W|$ are the lengths of V and W respectively.

Proof. We have to delete $|V| - s(V,W)$ terms from V and insert $|W| - s(V,W)$ terms to transform V into W as the best strategy. Moreover, we can edit V into W by the following idea.



V = ATCTGAT

So, we have to delete A, G, T in ATCTGAT and then insert G, A to obtain W. Now, the number of vertices not incident to an edge of the matching is $d(V,W)$. ∎

(*) So, it is sufficient to compute $s(V,W)$ in order to find the edit distance. (This is important in applications!)

Compute $s(V,W)$, $V = \langle v_1 v_2 \cdots v_n \rangle$ and $W = \langle w_1 w_2 \cdots w_m \rangle$

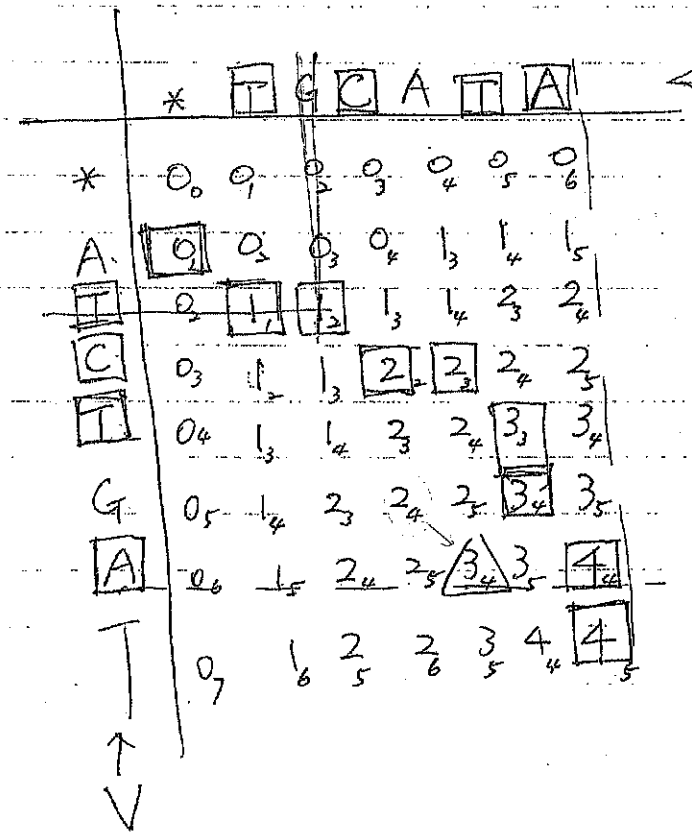
Algorithm (Dynamic Algorithm)

Define $s_{i,j} = s(V_i, W_j)$ where $V_i \equiv \langle v_1 v_2 \cdots v_i \rangle$ and $W_j = \langle w_1 w_2 \cdots w_j \rangle$, $s_{i,0} = s_{0,j} = 0$, $\forall\ 1 \le i \le n$ and $1 \le j \le m$.

Step 1

$$s_{i,j} = \max \begin{cases} s_{i-1,j} \\ s_{i,j-1} \\ s_{i-1,j-1} + 1 \end{cases} \quad \text{if } v_i = w_j.$$

One more common term.



$s(V,W) = 4$ $\quad d(V,W) = 5$ (edit distance)

$\checkmark\ d(V,W) = 7 + 6 - 2(4) = 5$. By (Fact 1).

Combinatorial Relations

Consider a permutation $\pi \in S_n$ (Symmetric group of order n)

(o) Find a longest increasing subsequence of π.

(oo) Find a longest increasing subsequence of π is equivalent to find the LCS between π and the permutation $(1\ 2\ 3\ \cdots\ n)$. $\begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ 1 & 2 & 3 & \cdots & n \end{pmatrix}$

→ identity permutation with cycle representation $(1)(2)(3)\cdots(n)$

(*) Every permutation $\pi = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ x_1 & x_2 & x_3 & \cdots & x_n \end{pmatrix}$ denoted by $\langle x_1\ x_2\ \cdots\ x_n \rangle$ has either an increasing or a decreasing subsequence of length at least $(\sim \sqrt{n})$.

Review (Discrete Mathematics)

Every sequence $\langle a_1, a_2, \cdots, a_{n^2+1} \rangle$ of real numbers contains either an increasing subsequence or a decreasing subsequence of length $n+1$.

Pf. Suppose that there is no increasing subsequence of length $n+1$. For $1 \le k \le n^2+1$, let the longest increasing

(以下是一个例子, $n=3$)

subsequence starting at a_k be m_k. By assumption, $m_k \le n$ for each $k \in \{1, 2, \cdots, n^2+1\}$. By Pingeon-hole principle, there exists an $1 \le n' \le n$, such that $m_{k_1} = m_{k_2} = \cdots = m_{k_{n+1}} = n'$. Now, consider $W =$ the subsequence $< a_{k_1}, a_{k_2}, \cdots, a_{k_{n+1}} >$. If $a_{k_i} \le a_{k_j}$, then m_{k_i} can be larger, a contradiction. Hence, W is a decreasing subsequence of length $n+1$. ∎

(前面加上 a_{k_i}, 这个数 a_{k_j}, ...)

=

Remark

Pigeon-hole principle plays an important role in showing the existence of certain configurations. Make sure you understand how to apply the principle.

$< 1, 7, 2, 3, 6, 5, 4, 10, 8, 9 >$

Sequence Comparison

$\pi = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ x_1 & x_2 & x_3 & \cdots & x_n \end{pmatrix} \in S_n$ (Symmetric group defined on $\{1, 2, 3, \cdots, n\}$)

π is denoted by $\langle x_1 x_2 \cdots x_n \rangle$.

Definition (Increasing subsequence of a permutation)

An increasing subsequence of a permutation $\pi = \langle x_1 x_2 \cdots x_n \rangle$ is a sequence of indices $1 \le i_1 \le i_2 \le \cdots \le i_k \le n$ s.t. $x_{i_1} \le x_{i_2} < \cdots \le x_{i_k}$.

(*) Finding the longest increasing subsequence (LIS) of a permutation $\pi = \langle x_1 x_2 \cdots x_n \rangle$ is equivalent to finding the longest common subsequence (LCS) of π and $\langle 1 2 3 \cdots n \rangle$.

(**) For sure, we can also use dynamic algorithm to find the answer. But, we shall use a non-dynamic programming approach to find an LIS in what follows.

(As finding $s(V, W)$.)

Young Tableaux

Definition (Partition of integer)

A partition of an integer n is a sequence of positive integers $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\ell$ s.t. $\sum_{i=1}^{\ell} \lambda_i = n$.

Notation:

$\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_\ell) \longmapsto \lambda \vdash n$ (reading: λ partitions n.)

or $(\lambda_1, \lambda_2, \cdots, \lambda_\ell) \vdash n$.

Definition (Young diagram of shape λ)

The Young diagram of shape λ is an array of n cells into ℓ left-justified rows with row i containing λ_i cells for $i = 1, 2, \cdots, \ell$.

(*) left-justified ↓ 後上而下

(*) right- 右的資料用由左而右

Examples:

1	2	5	8
4	7		
6			
9			

$(4,2,1,1)$

Not a Young tableau

("3 不見了!")

1	2	3	8
4	5		
6	7		
9			

$(4,2,2,1) \vdash 9$

standard

Young Tableau

Definition (Young Tableaux)

A Young tableau (of shape λ) is an array obtained by replacing the cells of the Young diagram λ with the numbers $1, 2, \cdots, n$ bijectively. A tableau is standard if its rows and columns are increasing sequences.

Permutation and (P,Q)-bitableau

Every permutation π of S_n can be represented by a pair of tableaux $(\lambda \vdash n)$, and there is a bijection between permutations and pairs of tableaux.

Note ① proposed first; Robinson 1938,

② In connection with representation theory, Schensted 1961

③ Knuth generalized the algorithm for the case of LCS, 1970

→ RSK algorithm

$$\pi \xrightarrow{RSK} (P, Q)$$

bijection

$\{$ Conjugacy classes of $S_n\}$

$\xleftrightarrow{1-1}$ $\{$ partitions of $n\}$

$\lambda \vdash n$

$\hookrightarrow \{$ Irreducible Representations of $S_n\}$

More about Representation Theory

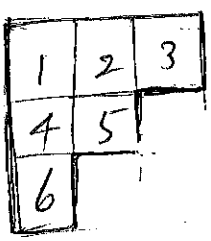
(*) Theorem Let d_λ be the number of distinct standard tableaux corresponding to $\lambda \vdash n$. Then

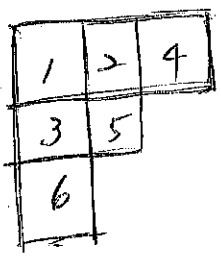
$$|S_n| = n! = \sum_{\lambda \vdash n} d_\lambda^2.$$

Note Tableaux 与 Tableau 的区别。

$n = 6$

$= 3 + 2 + 1$





We start with an example.

Example $\pi = \langle 7\ 2\ 8\ 1\ 3\ 4\ 10\ 6\ 9\ 5 \rangle \Leftrightarrow (P, Q)$

P: 7 → 2 7 ← insert 2 ↔ (上升序列長度1)

Q 在下面，依填入順序填入 1,2,…,10 在相同 shape 的 Tableau 中。

2 8 ← 接在第一列後面 ↔ (上升2)

1 8 ← insert 1 ↔ (上升4)

2

7

← insert 2

1 3 ← insert 3 ↔ (上升5)

2 8 ← 8比3大放右邊

7

1 3 4 ← 4放右邊 ↔ (上升3)

2 8

7

(註) 觀察第一列的數字個數即為 LIS.

1 3 4 10 ↔ (上升4)

2 8

7

10 → 下一page

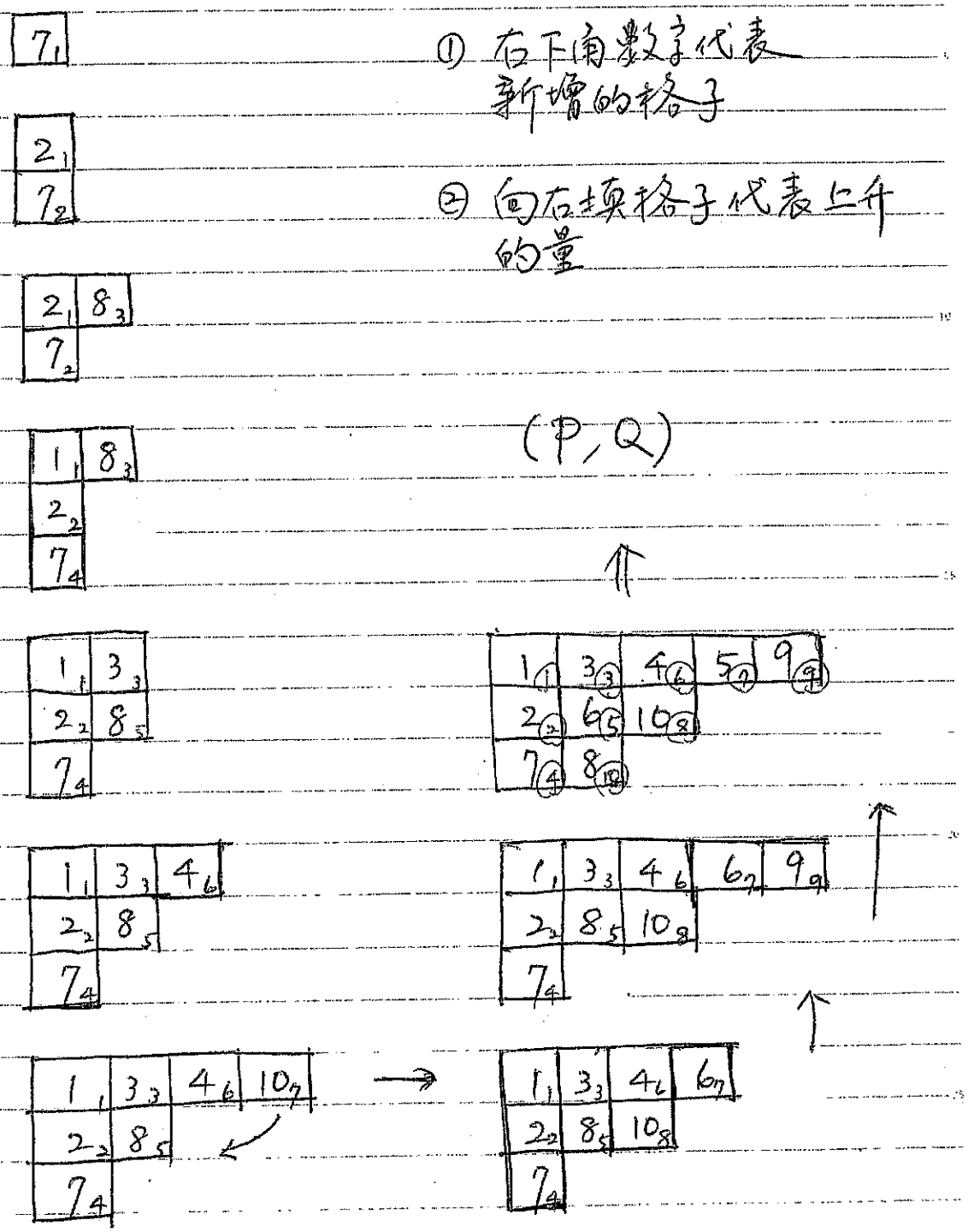
1 3 4 5 9 ↔ (上升5)

2 6 10

7 8

Step by Step

$\pi = \langle 7\ 2\ 8\ 1\ 3\ 4\ 10\ 6\ 9\ 5 \rangle$



Definition (Partial tableau)

A partial tableau P is a Young diagram with distinct entries whose rows and columns increase.

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 algorithm:

$R \leftarrow$ the first row of P

While x is less than some element of row R

Replace x_R^+ by x in R

$x \leftarrow x_R^+$

$R \leftarrow$ next row (down)

Place x at the end of row R.

(x is greater than every element of R)

Algorithm for finding (P, Q) from π.

(P, Q) will be obtained by a sequence of tableaux,

i.e. $(P_0, Q_0) = (\emptyset, \emptyset), (P_1, Q_1), (P_2, Q_2), \cdots, (P_n, Q_n) = (P, Q)$.

(Note 1) (Refer) Compare the following statements with example in next page.

(Note 2) (P_k, Q_k) is obtained from inserting x_k into P_k and put k in corresponding cell to make sure P_i's and Q_i's are of the same shape.

Starting from $(\emptyset, \emptyset)$, insert x_1 to obtain $(P_1, Q_1) =$ ($\boxed{x_1}$, $\boxed{1}$) respectively. Then, insert x_k to P_{k-1} to obtain P_k, and thus corresponding Q_k for $k = 2, \cdots, n$.

Permutation <4 2 3 6 5 1 7>, $n=7$

	P_i's	Q_i's	
$x_R^+ = 4$	4	1	(P_1, Q_1)
$2 \leftarrow 4$ ↓	2 / 4	1 / 2	(P_2, Q_2)
$3 > 2$ ($x=3$ 比第一列的數大)	2 3 / 4	1 3 / 2	(P_3, Q_3)
$6>2$ and $6>3$	2 3 6 / 4	1 3 4 / 2	(P_4, Q_4)
5 取代 6, 6 到下一列.	2 3 5 / 4 6	1 3 4 / 2 5	(P_5, Q_5)
$x_R^+ = 2$ $1 \leftarrow 2$ ↓下一列 $x_R^+ = 4$ $2 \leftarrow 4$ ↓下一列	1 3 5 / 2 6 / 4	1 3 4 / 2 5 / 6	(P_6, Q_6)
$7 > 1, 3, 5$	1 3 5 7 / 2 6 / 4	1 3 4 7 / 2 5 / 6	$(P_7, Q_7) = (P, Q)$

How to return to <4 2 3 6 5 1 7> from knowing (P, Q)?

① Use (P_7, Q_7) find $x_7 = 7$ since 7 is in the 7th position in Q_7 and it is in the 1st row.

② Delete 7, we have (in P_7 and Q_7)

1	3	5
2	6	
4		

1	3	4
2	5	
6		

$x = 4$, 往上一列而"2"小於4且為最大,

(用 4 取代 2; $x = 2$ 再往上一列而用 2 取代 1)

③ Delete "1",

2	3	5
4	6	

1	3	4
2	5	

, $x_6 = 1$

④ $x = 6$, 往上一列 5 是 x_R, 用 5 取代 6, $x_5 = 5$

⑤

2	3	6
4		

1	3	4
2		

$x_4 = 6$

⑥

2	3
4	

1	3
2	

$x_3 = 3$

⑦

2
4

1
2

$x = 4$, 往上一列
$2 < 4$, 用 2 取代 4,
$x_2 = 2$

$x_1 = 4$

Theorem The map $\pi \xrightarrow{RSK} (P,Q)$ is a bijection between elements of S_n and pairs of standard tableaux of the same shape $\lambda \vdash n$

Proof. It suffices to prove that a pair of tableaux (P,Q) for $\lambda \vdash n$ can determine a unique permutation. We shall obtain this permutation starting from finding x_n and then $x_{n-1}, x_{n-2}, \cdots, x_1$

Assume that (P_k, Q_k) has been constructed, $k = n, n-1, \cdots, 2$.

Then, we will find (P_{k-1}, Q_{k-1}) and x_k. First, we find the cell (i,j) containing k in Q_k. (the largest number)

$P_k(i,j)$ must have been the last element to be displaced in the construction of P_k, and thus "x_k" can be found. Now, (How?)

we shall use the following procedure to delete $P_k(i,j) = x_k$ from P_k. This gives P_{k-1} and Q_{k-1} can be obtained accordingly

Set $x \leftarrow P_k(i,j)$ and erase $P_k(i,j)$

(*) 第二列的上一列第"0"列 $R \leftarrow$ the $(i-1)$-st row of P_k

While R is not the zeroth row of P_k

(**) x_R 為在 x 的最大數 (R中)

Replace x_R by x in R

(***) 可以比上述移動 $R \leftarrow$ next row up

$x_k \leftarrow x$

∎

Lemma If $\pi = \langle x_1 x_2 x_3 \cdots x_n \rangle$ and x_k enters P_{k-1} (to become P_k) in column j, then the longest increasing subseq. of π ending in x_k has length j.

Proof. By induction on k. Clearly, it is true for $k=1$. So, suppose it holds for all values up to $k-1$. First, we claim, indeed, there exists an increasing subsequence of length j ending in x_k. Remind that x_k enters P_{k-1} in column j. Let y be the element of P_{k-1} in cell $(1, j-1)$. Then, $y < x_k$ since x_k enters in column j. By induction, there is an increasing subseq. ending in y with length $j-1$, thus we have an increasing subseq. of length j ending in x_k.

Now, we prove that there cannot be a longer increasing subseq. ending in x_k. If there exists one, let x_i be an (length $> j$) element preceding x_k. By induction, x_i enters (earlier) to the right of column j (in this increasing subsequence). Now, the element in the cell $(1, j)$, z, of P_i satisfies $z \leq x_i < x_k$. So, in order to enter column j, $x_k < z$, a contradiction. (By RSK algorithm.) ∎

Theorem Let $P(\pi)$ be the standard tableau of π. Then, the length of the longest increasing subseq. of π is the length of the first row of $P(\pi)$.

Proof. $P(\pi)$ is a standard tableau. (The first row has largest length.)
(maximum)

$\square$

Average Length of Longest Common Subsequence

(*) Let $\mathcal{V}$ and $\mathcal{W}$ be two sets of sequences of length n.

(*) Let $p : \mathcal{V} \times \mathcal{W} \to \mathbb{R}$ be a measure. For example, let $p : \mathcal{V} \times \mathcal{W} \to \left\{ \frac{1}{|\mathcal{V}| \cdot |\mathcal{W}|} \right\}$ a constant mapping.

(*) The average length of LCS between $\mathcal{V}$ and $\mathcal{W}$ is

$$a(n) = \frac{1}{|\mathcal{V}| |\mathcal{W}|} \cdot \sum_{\substack{V \in \mathcal{V}, \\ W \in \mathcal{W}}} a(V, W).$$

(**) Average longest increasing subseq. in random permutation in S_n,

$$a_{per}(n) = \frac{1}{1 \cdot n!} \sum_{\pi \in S_n} l.i.s.(\pi)$$ where $l.i.s.(\pi)$ is the length of the longest increasing subsequence of π.

Date / /

No. 17

(•) The problem of finding $s_{per}(n)$ was raised by Ulam, 1961; and Hammersley in 1972 proved that

$$\lim_{n\to\infty} \frac{s_{per}(n)}{\sqrt{n}} = s_{per} \text{ (a constant)}.$$

He also proved that $\frac{\pi}{2} \le s_{per} \le e$. (Surprisingly!)

Kingman, 1973, $1.59 < s_{per} < 2.49$.

Logan and Shepp, 1977; and Vershik and Kerov, 1977,

$$\boxed{s_{per} = 2.}$$

How about Longest Common Subsequence $s_k(n)$ in k-letter alphabet. (DNA sequences : $k=4$).

So, both $\mathcal{V}$ and $\mathcal{W}$ contain all k^n n-letter words in k-letter alphabet, $p(V,W) = \frac{1}{k^n \cdot k^n}$ for each $V \in \mathcal{V}$ and $W \in \mathcal{W}$.

Chvatal and Sankoff, 1975 : $\lim_{n\to\infty} \frac{s_k(n)}{n} = s_k$ (a constant)

Conjectures (1) $\lim_{k\to\infty} (s_k \cdot \sqrt{k}) = 2$.

(2) $s_k = \frac{2}{1+\sqrt{k}}$. (??)