Asymptotics in Sequences Comparisons

J.C. Breton

Ü Işlak

IESC, 15/05/2018.

Longest Common Subsequence

$n = 15$

$m = 3$

Longest Common Subsequence

D	A	B	A		D	A	B		D	D	B			A		A	D	B	D
D	A		A	B	D	A		A	D	D	B	D	B	A	D	A			

D	A		B	D-	. _ _ . .
D	A	A	B	D	. . . .

Length of the Longest Common Subsequence

$LC_n = 10$.

Longest common subsequence

= alignment (optimal) with holes

Algorithm (Dynamic Programming)

$$|LCS(x_1 \cdots x_{\ell_2}; y_1 \cdots y_\ell)|$$

$$= \begin{cases} |LCS(x_1 \cdots x_{\ell_2 - 1}; y_1 \cdots y_{\ell-1})| + 1 & \text{if } x_{\ell_2} = y_\ell, \\ \max\left(|LCS(x_1 \cdots x_{\ell_2}; y_1 \cdots y_{\ell-1})|, |LCS(x_1 \cdots x_{\ell_2 - 1}; y_1 \cdots y_\ell)|\right) & \\ \quad \text{if } x_{\ell_2} \neq y_\ell. \end{cases}$$

$$|LCS(x_1 \cdots x_{\ell_2}; 0)| = |LCS(0; y_1 \cdots y_\ell)| = 0$$

$$\forall\ \ell_2 = 1, 2, \ldots, n.$$
$$\forall\ \ell = 1, 2, \ldots, n$$

Edit/Levenstein distance

$$d(x_1 \cdots x_n; y_1 \cdots y_n) = 2(n - LC_n)$$

Penalties

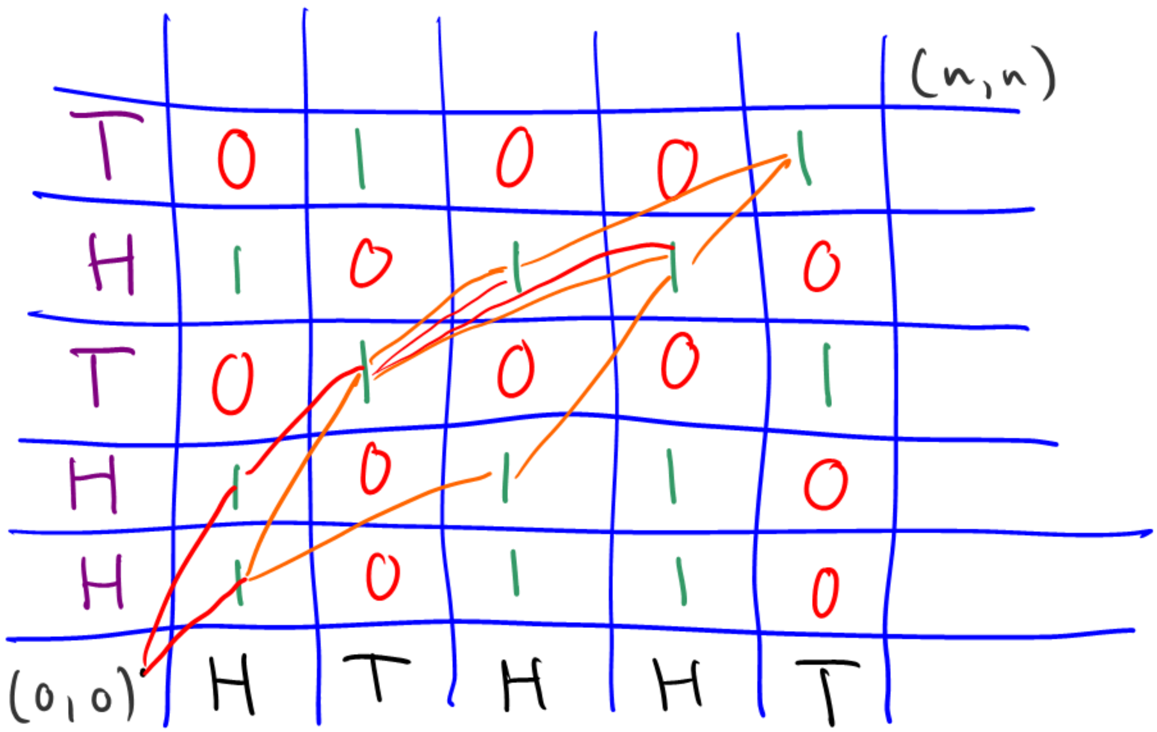
christian
krystyaan

c		h	r	i		s	t	i	a	a	n
	k		r		y	s	t	y	a		n

$X_1 \cdots X_n =$ H T H H T

$Y_1 \cdots Y_n =$ H H T H T

6 Common Subsequences, $LC_n = 4$.



LC_n = length of the longest strictly increasing north-east path from $(0,0)$ to (n,n).

$X_1, X_2, \ldots \ldots, X_n, \ldots$ iid with values in

$$\mathcal{A}_m = \{\alpha_1, \alpha_2, \ldots \ldots, \alpha_m\}$$

$Y_1, Y_2, \ldots \ldots, Y_n, \ldots$ iid with values in

$$\mathcal{A}_m = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$$

$$X = (X_i)_{i \geq 1} \perp\!\!\!\perp Y = (Y_i)_{i \geq 1}$$

LC_n = Length of the longest Common subsequence of

$X_1 \ldots X_n$ and $Y_1 \ldots Y_n$.

$$LC_n = \max_{k=1,\ldots,n} k : \exists \; \begin{array}{l} 1 \leq i_1 < i_2 < \cdots < i_k \leq n \\ 1 \leq j_1 < j_2 < \cdots < j_k \leq n \end{array} \quad \text{t.q}$$

$$X_{i_s} = Y_{j_s}, \; s = 1, 2, \ldots, k.$$

$\mathbb{E} LC_n \sim ?$, $\operatorname{Var} LC_n \sim ?$, $\dfrac{LC_n - \mathbb{E} C_n}{(\operatorname{Var} LC_n)^{\gamma}} \Longrightarrow ?$

Asymptotic behavior in Mean, Variance, Distribution of LC_n ?

Chvátal-Sankoff (1975)

$$LCS(X_1 \cdots X_{n_1+n_2}; Y_1 \cdots Y_{n_1+n_2}) \geq LCS(X_1 \cdots X_{n_1}; Y_1 \cdots Y_{n_1}) + LCS(X_{n_1+1} \cdots X_{n_1+n_2}; Y_{n_1+1} \cdots Y_{n_1+n_2})$$

So taking expectation and iid (stationarity)

$\mathbb{E} LC_{n_1+n_2} \geq \mathbb{E} LC_{n_1} + \mathbb{E} LC_{n_2}$. So

Fekete's Lemma, $\frac{\mathbb{E} LC_n}{n} \xrightarrow{n \to +\infty} \gamma_m^* = \sup_{k \geq 1} \frac{\mathbb{E} LC_k}{k}$

$\gamma_2^* \simeq 0.812$, $\gamma_3^* \simeq 0.717$, $\gamma_4^* \simeq 0.654$, ...

(iid uniform case).

Kiwi, Loebl, Matoušek (2005),

$$\lim_{m \to +\infty} \sqrt{m}\, \gamma_m^* = 2.$$

(Conjecture of Sankoff and Mainville, 1983)

Links with Ulam's Problem.

Speed of convergence

$$n\gamma_m^* - K_A\sqrt{n\log n} \le \mathbb{E}\, LC_n \le n\gamma_m^* \quad \text{(Alexander 1994)}$$

Variance ?

Efron-Stein Inequality / Tensorisation

$X_1, X_2, \dots, X_n$ $\perp\!\!\!\perp$, $X_i \sim \mu_i$, $\mu^n = \bigotimes_{i=1}^n \mu_i$

$f : \mathbb{R}^n \longrightarrow \mathbb{R}$,

$$\mathrm{Var}\, f(X_1, \dots, X_n) \le \mathbb{E} \sum_{i=1}^n \mathrm{Var}_{\mu_i} f(X_1, \dots, X_n)$$

$$= \frac{1}{2}\mathbb{E}\sum_{i=1}^n \mathbb{E}\underbrace{\left(f(X_1,\dots,X_n) - f(X_1,\dots,X_{i-1},\hat{X}_i,X_{i+1},\dots,X_n)\right)^2}_{\le 1}$$

$$LC_n = f(X_1,\dots,X_n; Y_1,\dots,Y_n)$$

Hence, $\mathrm{Var}\, LC_n \le \frac{1}{2} 2n = n$ (Steele 1986)

$\mathrm{Var}\, LC_n \geq Kn$? Yes in some biased cases (Lember, Matzinger, Amsalu, Gong, Ma, C.H., ...) IID Uniform ?

Longest Increasing Subsequences in Random Words

Order $1 = D, \; 2 = B, \; 3 = A$

1 3 2 3 1 3 2 1 1 2 3 3 1 2 1

$n = 15$

$m = 3$

Instead of two sequences, just one:

Longest Increasing Subsequence

1 1 1 1 2 3 3

Alphabet $\mathcal{A}_5 : \{ \alpha_1 < \alpha_2 < \alpha_3 < \alpha_4 < \alpha_5 \}$ $\quad m = 5$

Wd : $\alpha_1 \alpha_3 \alpha_2 \alpha_4 \alpha_2 \alpha_3 \alpha_5 \alpha_4 \alpha_2 \alpha_4 \alpha_3 \alpha_5 \alpha_4$ $\quad n = 12$

LI_n = length of the longest increasing subsequence

$\alpha_1 \alpha_2 \alpha_2 \alpha_3 \alpha_4 \alpha_4 \alpha_5$

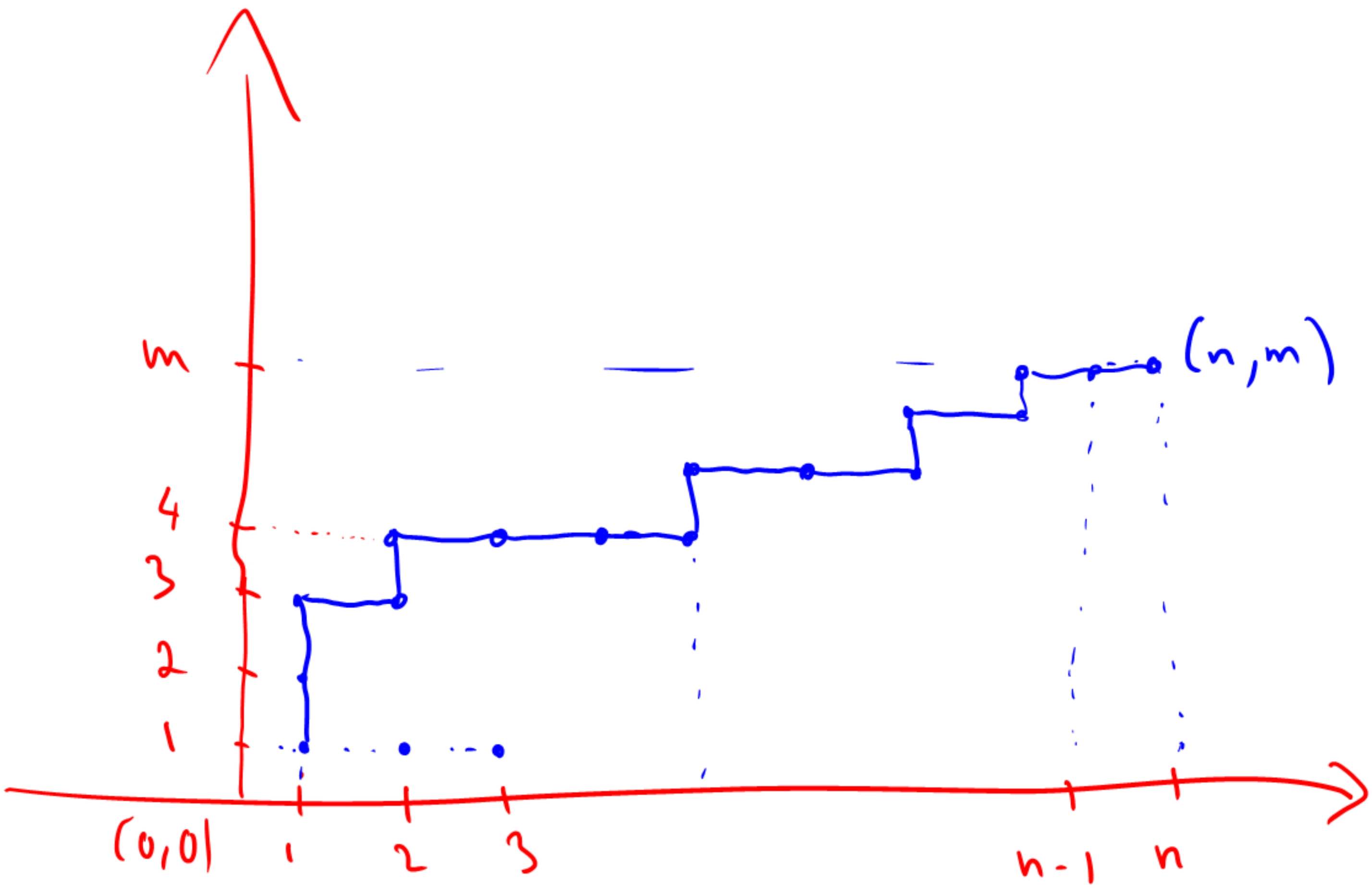
$\alpha_1 \alpha_2 \alpha_2 \alpha_3 \alpha_4 \alpha_4 \alpha_4$

$X_1, X_2, \ldots, X_n, \ldots$ iid $\quad \mathcal{A}_m : \{ \alpha_1 < \alpha_2 < \cdots < \alpha_m \}$

$LI_n = \max k : \exists \; 1 \le i_1 < i_2 < \cdots < i_k \le n$

s.t. $X_{i_1} \le X_{i_2} \le \cdots \le X_{i_k}$

LI_n and Last Passage Percolation



$$LI_n = \max_{\pi \in NE} \sum_{(i,j)\in\pi} \mathbb{1}_{\{X_i = \alpha_j\}}$$

NE = North-East paths from $(1,1) \nearrow (m,n)$

The Bernoulli r.v. $\mathbb{1}_{\{X_i=\alpha_j\}}$ are dependent! (Exchangeable in the uniform case).

<u>Kerov (1994)</u> Let $\mathbb{P}(X_1 = \alpha_k) = \frac{1}{m}$,

$k = 1, \dots, m$.

$$\frac{LI_n - n/m}{\sqrt{n/m}} \underset{n \to +\infty}{\Longrightarrow} \lambda^0_{max}, \text{ where } \lambda^0_{max}$$

is the maximal eigenvalue of the traceless $m \times m$ GUE, i.e.,

$M - \frac{1}{m}(\mathrm{Tr}\, M) I$, where $M \in GUE(m \times m)$.

* Tracy-Widom (2001), Johansson (2001) *

Baryshnikov (2001), Gravner, T-W (2001)

$$\lambda_{max} \stackrel{\mathcal{L}}{=} \max_{0 = t_0 \le t_1 \le \cdots \le t_{m-1} \le t_m = 1} \sum_{i=1}^{m} \left(B^{(i)}(t_i) - B^{(i)}(t_{i-1}) \right).$$

Doumerc, O'Connell, Yor, Bougerol, Jeulin

Longest Common and Increasing Subsequences

Order $1 = D, \ 2 = B, \ 3 = A$

$n = 15$

$m = 3$

1 3 2 3 1 3 2 1 1 2 3 3 1 2 1

1 3 3 2 1 3 3 1 1 2 1 2 3 1 3

Length of the longest common and increasing subsequence

$$LCI_n = 7$$

$(X_i)_{i \geq 1}, \ (Y_i)_{i \geq 1},$

$$LCI_n = \max_{k=1,\dots,n} k : \exists \ 1 \leq i_1 < i_2 < \dots < i_k \leq n, \quad 1 \leq j_1 < j_2 < \dots < j_k \leq n$$

s.t. $X_{i_s} = Y_{j_s}, \quad s = 1, 2, \dots, k$

$$X_{i_1} \leq X_{i_2} \leq \dots \leq X_{i_k}$$

$$Y_{i_1} \leq Y_{i_2} \leq \dots \leq Y_{i_k}$$

Asymptotic Behavior of LCI_n ?

Théorème 1 (J.-C. Breton, C.H.). Let $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$ be two (independent) sequences of iid random variables uniformly distributed on $\mathcal{A}_m = \{\alpha_1 < \alpha_2 < \cdots < \alpha_m\}$. Then,

$$\frac{LCI_n - n/m}{\sqrt{n/m}} \underset{n\to+\infty}{\Longrightarrow} \max_{0=t_0\leq t_1\leq\cdots\leq t_{m-1}\leq t_m=1} \min\left(-\frac{1}{m}\sum_{i=1}^m B_1^{(i)}(1) + \sum_{i=1}^m \left(B_1^{(i)}(t_i) - B_1^{(i)}(t_{i-1})\right), -\frac{1}{m}\sum_{i=1}^m B_2^{(i)}(1) + \sum_{i=1}^m \left(B_2^{(i)}(t_i) - B_2^{(i)}(t_{i-1})\right)\right)$$

where

$$B_1 = \left(B_1^{(1)}, \ldots, B_1^{(m)}\right) \quad \text{et} \quad B_2 = \left(B_2^{(1)}, \ldots, B_2^{(m)}\right)$$

are two m-dimensional standard BM (and $\perp\!\!\!\perp$).

Random Matrix Interpretation ??

Particular cases

- $X = Y$, $B_1 = B_2$, $\cancel{m \times n}$, $LCI_n = LI_n$.

$$\frac{LI_n - n/m}{\sqrt{n/m}} \underset{n \to +\infty}{\Longrightarrow} \max \sum_{i=1}^{m} \left(B^{(i)}(t_i) - B^{(i)}(t_{i-1})\right) - \frac{1}{m}\sum_{i=1}^{m} B^{(i)}(1).$$

$$\underbrace{t_0 = 0 \le t_1 \le \cdots \le t_{m-1} \le t_m = 1}_{\text{Weyl Chamber}}$$

Above result is very natural.

$$LI_n = \#\alpha_1 + \#\alpha_2 + \cdots + \#\alpha_m.$$

$$= \max_{0 = k_0 \le k_1 \le k_2 \cdots \le k_{m-1} \le k_m = n} \left(\#\alpha_1 + \#\alpha_2 \text{ after } k_1 + \cdots + \#\alpha_m \text{ after } k_{m-1} \right)$$

Moreover $\#\alpha_1 + \#\alpha_2 + \cdots + \#\alpha_m = n$!

But $\#\alpha_i$ after k_{i-1} and before k_i $= \#$ of α_i before k_i $- \#$ of α_i before k_{i-1}.

Then

$$LI_n - \frac{n}{m} = -\frac{1}{m}\sum_{i=1}^{m} S_n^i + \max\left(S^1_{k_1} + S^2_{k_2} + \cdots + S^{m-1}_{k_{m-1}}\right) + o(\sqrt{n})$$

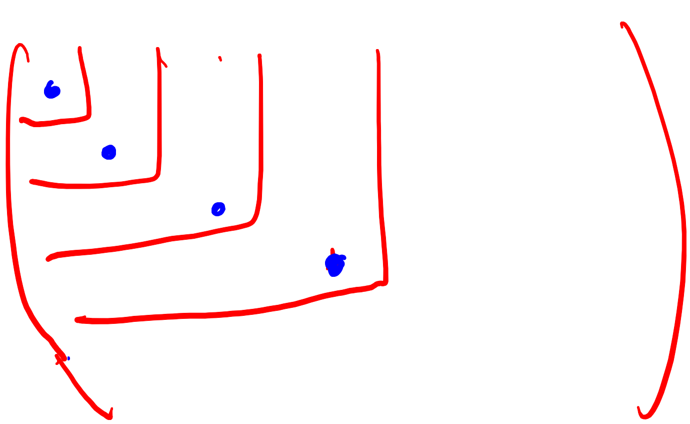
where $(S^{i}_{k_i})$ are random walks.

Hence Donsker's

$$\frac{L I_n - n/m}{\sqrt{n/m}} \Rightarrow \underbrace{-\frac{1}{m}\sum_{i=1}^{m} B^{(i)}(1) + \underbrace{\max_{0=t_0\le t_1\le \cdots \le t_{m-1}\le t_m=1} \sum_{i=1}^{m}\left(B^{(i)}(t_i)-B^{(i)}(t_{i-1})\right)}_{\stackrel{d}{=}\ \lambda_{\max}(GUE_{m\times m})}}_{\stackrel{d}{=}\ ?}$$

$$-\frac{1}{m}\sum_{i=1}^{m}\lambda_i + \lambda_{\max}$$

YES

$M =$ 

Benaych-Georges + C.H.

• In general, the dependence structure of X with Y is reflected into a dependence structure between the two m-dimensional BM, B_1 and B_2.

• $m = 2$, X and Y $\perp\!\!\!\perp$, then

$$\frac{LCI_n - n/2}{\sqrt{n}} \Longrightarrow \max_{0 \le t \le 1} \min\left(B_1(t) - \frac{1}{2} B_1(1), B_2(t) - \frac{1}{2} B_2(1)\right),$$

where B_1 and B_2 are two $\perp\!\!\!\perp$ standard linear B.M.

Lember, Matzinger, C.H. (2006).

Proof: "LCI_n"

(A) Combinatorial Representation of LCI_n.

LCI_n = max over random constraints of min of random sums of randomly stopped random variables.

(B) Derandomization. $\mathbb{E}$.

(C) Weak Invariance Principles (Donsker).

Back to LC_n (permutations/Words)

Ulam's Problem.

S_n : symmetric group, $\sigma \in S_n$, $LI_n(\sigma)$, the length of the longest $\uparrow$ subsequence of σ. Study of the r.v. $LI_n(\sigma)$, when S_n is endowed with the uniform measure.

Ulam: $\mathbb{E}\, LI_n(\sigma) \simeq 1.7\sqrt{n}$

$\operatorname{Var} LI_n(\sigma) \simeq$

$$\frac{LI_n(\sigma) - \mathbb{E}\, LI_n(\sigma)}{\sqrt{\operatorname{Var} LI_n(\sigma)}} \Longrightarrow N(0,1).$$

(Monte-Carlo Calculations in Problems of Mathematical Physics). Late 50's - early 60's.

Vershik and Kerov (1977), Logan and Shepp (1977)

$$\frac{E\, LI_n(\sigma)}{\sqrt{n}} \longrightarrow 2.$$

Conjecture (Simulations / Percolations) $\simeq$ 1980's

$$\operatorname{Var} LI_n(\sigma) \approx n^{1/3}.$$ Odlyzko + Rains, Kesten

Baik - Deift - Johansson (1999)

$$\frac{LI_n(\sigma) - 2\sqrt{n}}{n^{1/6}} \Longrightarrow F_2.$$

F_2 : Tracy-Widom law, limiting law of max eigenvalue of $n \times n$ GUE.

F_2 also appears as limiting law of LC_n of two random permutations? (Aldous - Diaconis $\simeq$ 2000).

Proposition (Easy). Let σ and π be two $\perp\!\!\!\perp$ uniform random permutations in $\mathfrak{S}_n$. Let $LC_n(\sigma,\pi)$ be the length of the longest common subsequence of $(\sigma_1, \dots, \sigma_n)$ and $(\pi_1, \dots, \pi_n)$. Then,

$$\frac{LC_n(\sigma,\pi) - 2\sqrt{n}}{n^{1/6}} \Longrightarrow F_2 .$$

Theorem 2 (Hard). Let $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$ be two independent sequences of iid random variables with values in a finite alphabet $\mathcal{A}_m$. Let LC_n be the length of the longest common subsequence of $X_1 \dots X_n$ and $Y_1 \dots Y_n$. If $\operatorname{Var} LC_n \geq Kn$, then

$$\frac{LC_n - \mathbb{E} LC_n}{\sqrt{\operatorname{Var} LC_n}} \Longrightarrow N(0,1).$$

Proof of Proposition

Fact 1. Note that $LI_n(\sigma)$, the length of the longest increasing subsequence of $\sigma \in S_n$ is equal to

$$LC_n((1,\dots,n),(\sigma_1,\dots,\sigma_n))\,!$$

$$\begin{array}{lllll} id: & 1 & 2 & \dots & n \\ \sigma: & \sigma_1 & \sigma_2 & \dots & \sigma_n \end{array}$$

Fact 2. Let $a=(a_1,\dots,a_n)$ be a fixed permutation of $1,\dots,n$. Let σ be a uniform random permutation in S_n. Then,

$$LI_n(\sigma) \stackrel{d}{=} LC_n\big((a_1,\dots,a_n);(\sigma_1,\dots,\sigma_n)\big)$$

Proof. Let $\rho \in S_n$ s.t. $\rho(i)=a_i$, $\forall i=1,\dots,n$.

Then clearly,

$\tilde{\sigma} = \sigma\rho$ is a uniform random permutation in S_n.

$$LC_n((a_1,\dots,a_n);(\sigma_1,\dots,\sigma_n))$$

$$= LC_n((\rho(1),\dots,\rho(n)),(\sigma_1,\dots,\sigma_n))$$

$$\overset{d}{=} LC_n((\rho(1),\dots,\rho(n)),(\tilde{\sigma}_1,\dots,\tilde{\sigma}_n))$$

$$= LC_n((\rho(1),\dots,\rho(n)),(\sigma_{\rho(1)},\dots,\sigma_{\rho(n)}))$$

$$\overset{d}{=} LC_n((1,\dots,n),(\sigma_1,\dots,\sigma_n)).$$

$$= LI_n(\sigma), \text{ by Fact 1.}$$

<u>Fact 3</u>. Let σ and π be two $\perp\!\!\!\perp$ uniform random permutations in S_n. Then,

$$\mathbb{P}(LC_n(\pi,\sigma) \leq x) = \mathbb{P}(LI_n(\sigma) \leq x)$$

<u>Proof</u>. Let $\mathcal{A} := \{a = (a_1,\dots,a_n) : a_i \in \{1,\dots,n\}$ and $a_i \neq a_j$ for $i \neq j\}$

$$\mathbb{P}(LC_n(\pi,\sigma) \leq x)$$

$$= \sum_{a \in \mathfrak{S}} \mathbb{P}\big(LC_n(\pi, \sigma) \le x \mid (\pi_1, \dots, \pi_n) = a\big)\, \mathbb{P}\big((\pi_1, \dots, \pi_n) = a\big)$$

$$= \frac{1}{n!} \sum_{a \in \mathfrak{S}} \mathbb{P}\big(LC_n((a_1, \dots, a_n); (\sigma_1, \dots, \sigma_n)) \le x\big)$$

Fact 2

$$\overset{\downarrow}{=} \frac{1}{n!} \sum_{a \in \mathfrak{S}} \mathbb{P}\big(LI_n(\sigma) \le x\big)$$

$$= \mathbb{P}\big(LI_n(\sigma) \le x\big).$$

Hence, the Proposition follows (is equivalent) to the theorem of Baik - Deift and Johansson. $\blacksquare$

Proof of Theorem 2.

Three step method

a) Stein's Method (Chatterjee, Lachièze-Rey and Peccati)

b) Short String Genericity principle

c) Variance Estimates

Distances

Kolmogorov distance

$$d_K(\mu_1, \mu_2) = \sup_{x \in \mathbb{R}} \left| \mu_1(-\infty, x] - \mu_2(-\infty, x] \right|$$

$$= \sup_{h \in \mathcal{H}_1} \left| \int h \, d\mu_1 - \int h \, d\mu_2 \right|$$

$\mathcal{H}_1 = \{ \mathbb{1}_{(-\infty, x]} : x \in \mathbb{R} \}$

Monge-Kantorovich-Wasserstein distance

$$d_1(\mu_1, \mu_2) = \sup_{h \in \mathcal{H}_2} \left| \int h \, d\mu_1 - \int h \, d\mu_2 \right|$$

$$\mathcal{H}_2 = \{ h : \mathbb{R} \to \mathbb{R} : |h(x) - h(y)| \leq |x - y| \} = \mathrm{Lip}(1).$$

$$d_K(\mu_1, \mu_2) \leq \sqrt{2 C d_1(\mu_1, \mu_2)}, \text{ if}$$

μ_2 is a.c. with density bounded by C.

Chatterjee (2008) Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$,

let $W = (W_1, \dots, W_n)$ have $\perp\!\!\!\perp$ components

Let $0 < \sigma^2 = \operatorname{Var} f(W) < +\infty$. Then,

$$d_1\left(\frac{f(W) - \mathbb{E} f(W)}{\sqrt{\operatorname{Var} f(W)}}, Z\right) \leq \sqrt{\frac{\operatorname{Var} T}{\sigma^2}} + \frac{1}{2\sigma^3} \sum_{j=1}^{n} \mathbb{E}|\Delta_j f(W)|^3$$

Lachièze-Rey & Peccati (2016)

$$d_K\left(\frac{f(W) - \mathbb{E} f(W)}{\sqrt{\operatorname{Var} f(W)}}, Z\right) \leq \sqrt{\frac{\operatorname{Var} T}{\sigma^2}} + \sqrt{\frac{\operatorname{Var} \tilde{T}}{\sigma^2}}$$

$$+ \frac{1}{4\sigma^3} \sum_{j=1}^{n} \sqrt{\mathbb{E}|\Delta_j f(W)|^6} + \frac{\sqrt{2\pi}}{16\sigma^3} \sum_{j=1}^{n} \sqrt{\mathbb{E}|\Delta_j f(W)|^3},$$

where $\Delta_j f = \dots$, $T = \dots$

and $\tilde{T} = \dots$

Let W' be an $\perp\!\!\!\perp$ copy of W, then

$$\Delta_j f(W) = f(W) - f(W^j), \text{ where}$$

$$f(W^j) = f(W_1, W_2, \ldots, W_{j-1}, W_j', W_{j+1}, \ldots, W_n).$$

For any $A \subset \{1,\ldots,n\}$, let W^A be defined via $W_i^A = \begin{cases} W_i' & \text{if } i \in A \\ W_i & \text{if } i \notin A. \end{cases}$

Let $T_A = \sum_{j \notin A} \Delta_j f(W) \Delta_j f(W^A)$ and

$$T = \frac{1}{2} \sum_{A \subsetneq \{1,\ldots,n\}} \frac{T_A}{\binom{n}{|A|}(n-|A|)}.$$

$$\tilde{T}_A = \sum_{j \notin A} \Delta_j f(W) \left|\Delta_j f(W^A)\right|$$

$$\tilde{T} = \frac{1}{2} \sum_{A \subsetneq \{1,\ldots,n\}} \frac{\tilde{T}_A}{\binom{n}{|A|}(n-|A|)}$$

In our case, $W = (X_1, \dots, X_n; Y_1, \dots, Y_n)$ and

$$f(W) = LC_n(X_1 \cdots X_n; Y_1 \cdots Y_n).$$

But for $1 \leq j \leq n$,

$$|\Delta_j f(W)| = |LC_n(X_1 \cdots X_n; Y_1 \cdots Y_n) - LC_n(X_1 \cdots X_{j-1} X_j' X_{j+1} \cdots X_n; Y_1 \cdots Y_n)|$$

$$\leq 1,$$

and similarly for $n+1 \leq j \leq 2n$. So

$$\frac{1}{\sigma^3} \sum_{j=1}^{n} \mathbb{E}|\Delta_j f(W)|^3 = \frac{1}{\sigma^3} \sum_{j=1}^{2n} \mathbb{E}|\Delta_j LC_n|^3$$

$$\leq \frac{2n}{\sigma^3} = \frac{n}{\sigma^3} \leq \frac{1}{K^{3/2}\sqrt{n}},$$

if $\sigma^2 \geq Kn$. Similarly for $\mathbb{E}\sqrt{|\Delta_j LC_n|^6}$

How to estimate $\sqrt{\frac{\mathrm{Var}\, T}{\sigma^2}}$, $\sqrt{\frac{\mathrm{Var}\, T'}{\sigma^2}}$?

$$\operatorname{Var} T = \frac{1}{4} \operatorname{Var}\left(\sum_{A \subsetneq \{1,\dots,2n\}} \sum_{j \notin A} \frac{\Delta_j f(W) \Delta_j f(W^A)}{\binom{2n}{|A|}(2n-|A|)} \right)$$

$$= \frac{1}{4} \sum_{(A,B,j,k) \in S_1} \frac{\operatorname{Cov}\left(\Delta_j f(W) \Delta_j f(W^A), \Delta_k f(W) \Delta_k f(W^B)\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}$$

$$S_1 = \{ (A,B,j,k) : A \subsetneq \{1,\dots,2n\}, B \subsetneq \{1,\dots,2n\}, j \notin A, k \notin B \}.$$

General lemma, shows that

$$\operatorname{Var} T \leq \frac{n^2}{2}$$ (not good enough since $\sigma^2 \geq Kn$) need $o(n^2)$.

Hence $\frac{\sqrt{\operatorname{Var} T}}{\sigma^2}$ needs to be estimated differently

Replace S_1 by several subsets, estimate each parts ($\simeq 20$ pages) and use

Short Strings Genericity

(H. Matzinger and C.H.)

Assume $n = \nu d$, say ν is fixed, for now, (then small compared to n) and let the integers

$$r_0 = 0 \leq r_1 \leq r_2 \leq \dots \leq r_{d-1} \leq r_d = n$$

be such that

$$LC_n = \sum_{i=1}^{d} \left| LCS\left(X_{\nu(i-1)+1} \cdots X_{\nu i}; Y_{r_{i-1}+1} \cdots Y_{r_i}\right) \right|$$

where $LCS(\cdot\cdot ; \cdot\cdot)$ is the longest common subsequence of the words $X_{\nu(i-1)+1} \cdots X_{\nu i}$ and $Y_{r_{i-1}+1} \cdots Y_{r_i}$.

i.e., this is an optimal alignment aligning $[\nu(i-1)+1, \nu i]$ with $[r_{i-1}+1, r_i]$, for all $i = 1, 2, \dots, d$.

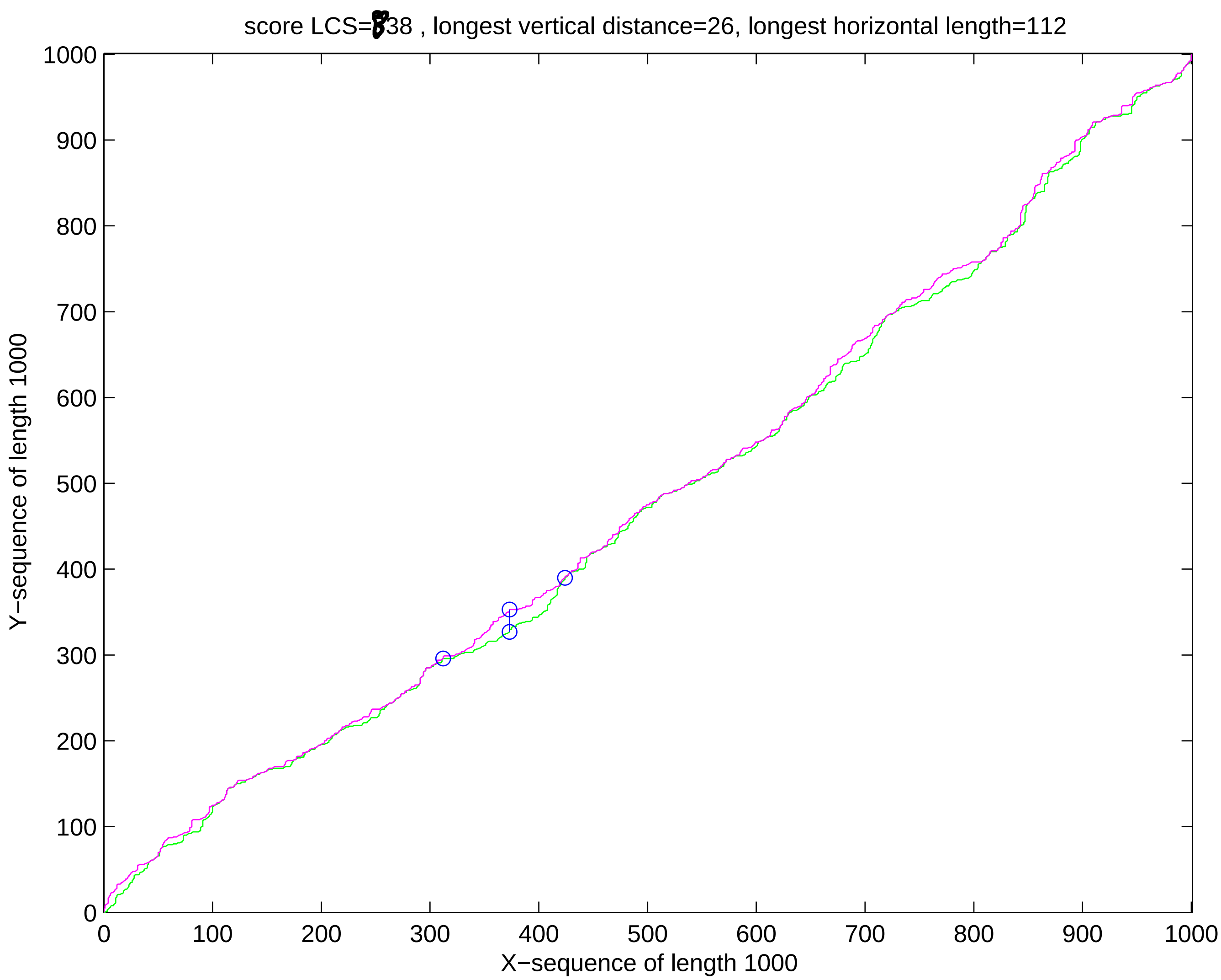
Such integers, $r_0, r_1, \dots, r_d$ always exists.

Genericity: Typically, the vast majority of the random lengths $r_i - r_{i-1}$ are close to v.

Quantitatively, take $v = n^\alpha$, $0 < \alpha < 1$, let E_n be the random sets of alignments with lengths $r_i - r_{i-1}$ close to v, then

$$\mathbb{P}(E_n) \geqslant 1 - e^{-n^{1-\alpha}(1+\log(1+n^\alpha))}.$$

When representing alignments in $\mathbb{N}^2$, we stay close to the diagonal at a macroscopic level.


score LCS=838 , longest vertical distance=26, longest horizontal length=112
Y−sequence of length 1000
X−sequence of length 1000
0
100
200
300
400
500
600
700
800
900
1000

To return to the estimation of $\operatorname{Var} T$, we condition again, this time with respect to E_n. When E_n holds only a "small number" of terms in the sums defining the variance. When E_n does not hold, then many terms in the sum but exponentially small probability that E_n does not occur.

$$\operatorname{Var} T \leq C\left(n^2 e^{-n^{1-\alpha}(1+\log(1+n^\alpha))} + n^{1+\alpha} + n^{3-\frac{3\alpha}{2}}\sqrt{\log n^\alpha}\right),$$

similarly for $\operatorname{Var} \tilde{T}$.

Take $\alpha = 4/5$, to get

$$d_K\left(\frac{LC_n - \mathbb{E}LC_n}{\sqrt{\mathrm{Var}\,LC_n}}, Z\right)$$

$$\leq C\sqrt{\frac{n^{9/5}}{\sigma^2}} + \frac{n}{\sigma^3}$$

But, $\sigma^2 \geq Kn$, so

$$\leq \frac{C}{K}\sqrt{\frac{1}{n^{1/5}}\log n} + \frac{1}{K^{3/2}\sqrt{n}}$$

$$\leq C\sqrt{\frac{1}{n^{1/5}}\log n} \xrightarrow[n\to+\infty]{} 0.$$

- Sublinear lower estimate on σ^2 is enough, $\sigma^2 \geq Kn^{9/10+\varepsilon}$ is enough
- All the cases where the variance is lower estimated, we have $\sigma^2 \geq Kn$!

Thank You!