Exact Asymptotic Results for Bernoulli Matching Model of Sequence Alignment

Sergei Nechaev and Satya Majumdar

LPTMS (Orsay, France)

Thanks to:

K. Mallick (*SPT, Saclay, France*)
M. Tamm (*Moscow University, Moscow*)

Search of common subsequence in two random sequences ("alignment problem")

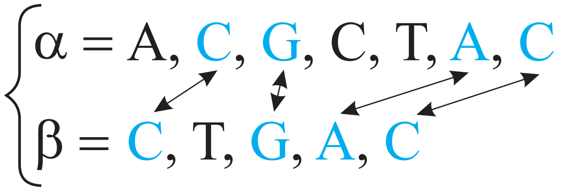
Example:

Consider two arbitrary sequences α and β. Each sequence is constructed from the 4-letter alphabet **A, C, G, T**:

$$\alpha = \{A, C, G, C, T, A, C\}$$

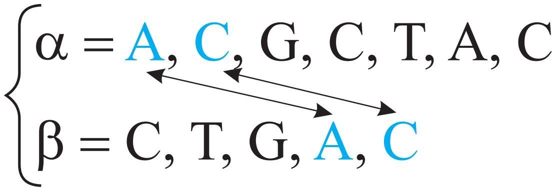
$$\beta = \{C, T, G, A, C\}$$

Common subword of two words α and β is their **ordered** common subsequence. For example, the subword {**C,G,A,C**} is a common subsequence of words α and β.



$$\begin{cases} \alpha = A, C, G, C, T, A, C \\ \beta = C, T, G, A, C \end{cases}$$

или



$$\begin{cases} \alpha = A, C, G, C, T, A, C \\ \beta = C, T, G, A, C \end{cases}$$

Total length=4

Total length=2

Principal question:

What is the statistics of the longest common subword?

The length $L_{i,j}$ of the Longest Common Subsequence (LCS) of two arbitrary words of lengths i and j can be computed in polynomial time $\sim O(ij)$ via the recursive algorithm

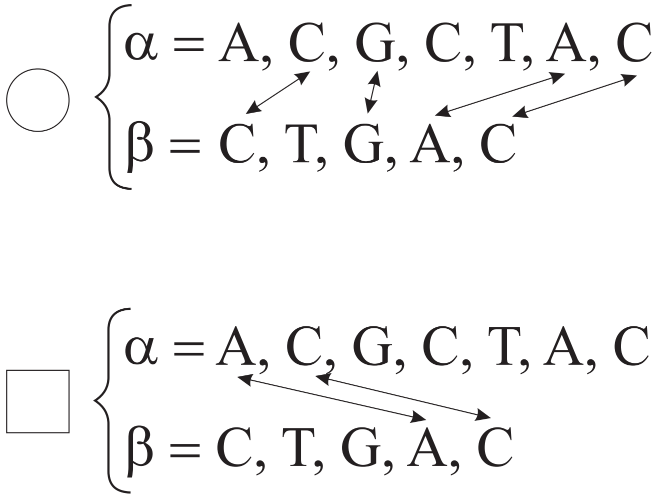
$$L_{i,j} = \max\left[L_{i-1,j}, L_{i,j-1}, L_{i-1,j-1} + \eta_{i,j} \right]$$

with the boundary conditions

$$L_{i,0} = L_{0,j} = L_{0,0} = 0$$

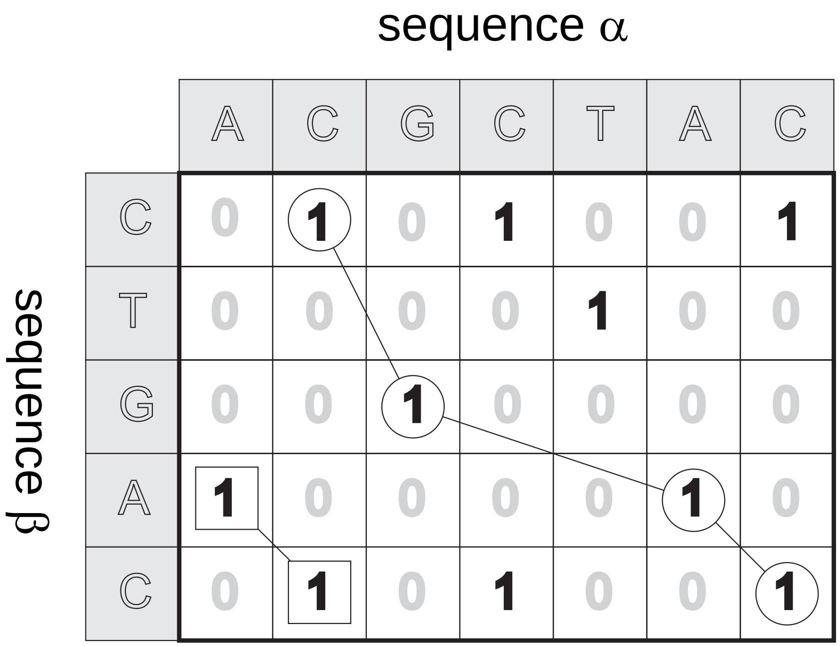
where the “noise” $\eta_{i,j}$ is defined as follows:

$$\eta_{i,j} = \begin{cases} 1 & \text{if the letters in the position } i \text{ (in } \alpha\text{) and } j \text{ (in } \beta\text{) are the same} \\ 0 & \text{otherwise} \end{cases}$$



○ $\begin{cases} \alpha = \mathrm{A, C, G, C, T, A, C} \\ \beta = \mathrm{C, T, G, A, C} \end{cases}$

□ $\begin{cases} \alpha = \mathrm{A, C, G, C, T, A, C} \\ \beta = \mathrm{C, T, G, A, C} \end{cases}$

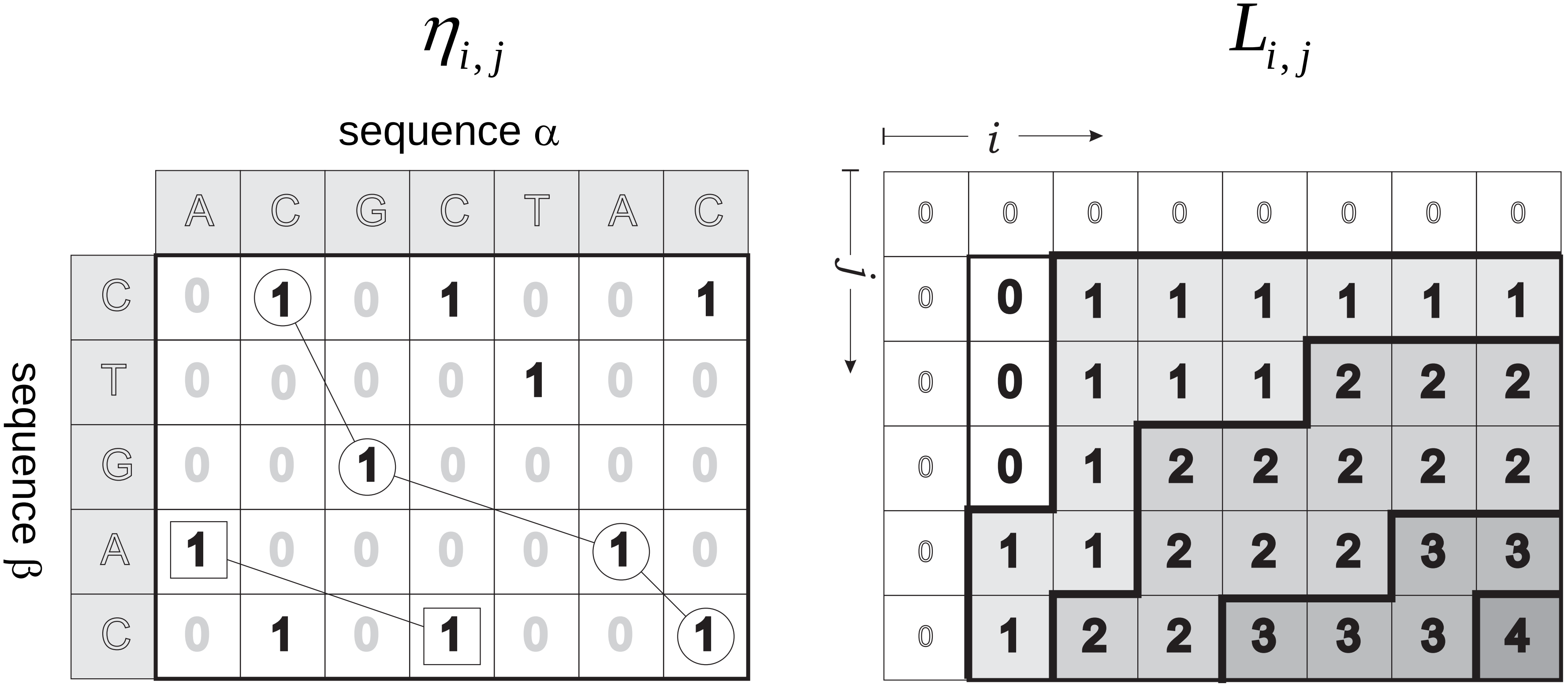


	A	C	G	C	T	A	C
C	0	1	0	1	0	0	1
T	0	0	0	0	1	0	0
G	0	0	1	0	0	0	0
A	1	0	0	0	0	1	0
C	0	1	0	1	0	0	1

Visualization of the recursive algorithm

$$L_{i,j} = \max\left[L_{i-1,j}, L_{i,j-1}, L_{i-1,j-1} + \eta_{i,j} \right]$$

$$\left(L_{i,0} = L_{0,j} = L_{0,0} = 0 \right)$$



$\eta_{i,j}$

	A	C	G	C	T	A	C
C	0	1	0	1	0	0	1
T	0	0	0	0	1	0	0
G	0	0	1	0	0	0	0
A	1	0	0	0	0	1	0
C	0	1	0	1	0	0	1

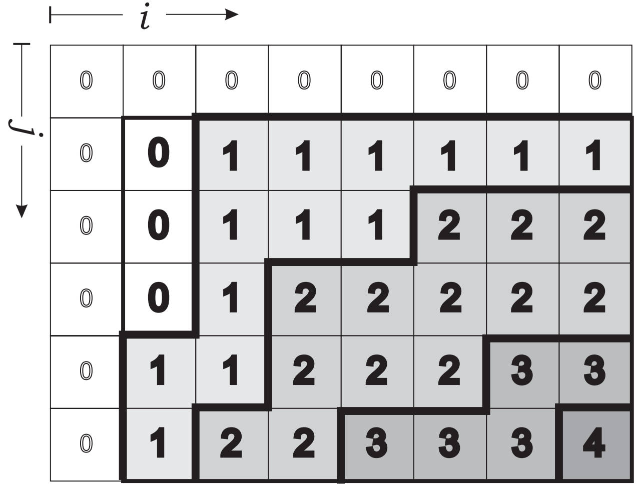
(a)

$L_{i,j}$

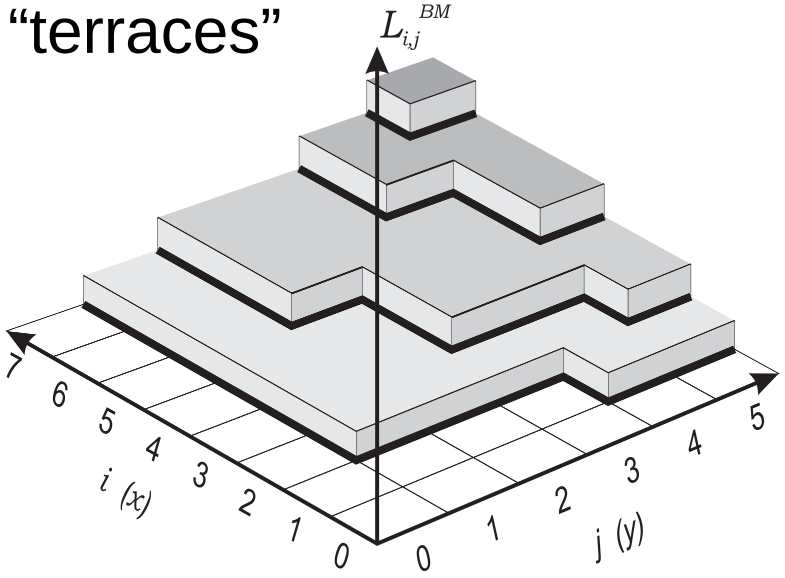
0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1
0	0	1	1	1	2	2	2
0	0	1	2	2	2	2	2
0	1	1	2	2	2	3	3
0	1	2	2	3	3	3	4

(b)

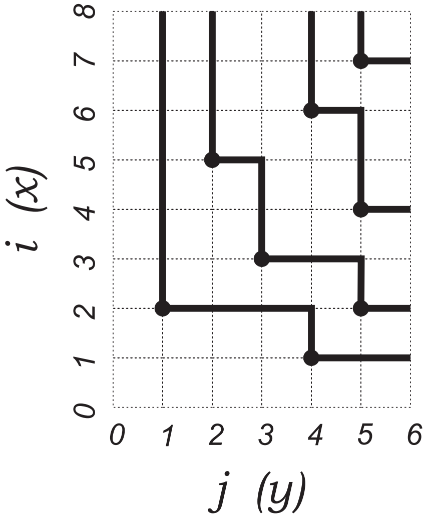
It is convenient to pass from the two-dimensional table $L_{i,j}$ to the (2+1)-dimensional representation in form of "terraces"

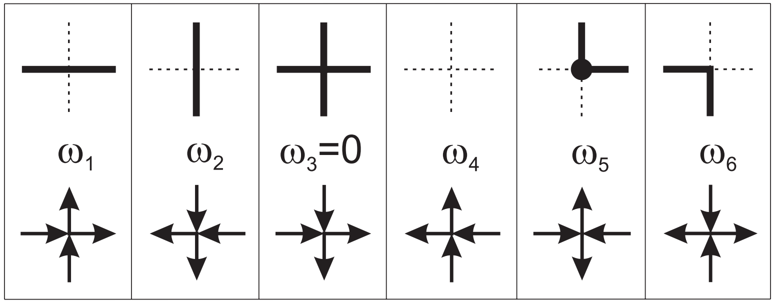


0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1
0	0	1	1	1	2	2	2
0	0	1	2	2	2	2	2
0	1	1	2	2	2	3	3
0	1	2	2	3	3	3	4



Connection to 5-vertex model





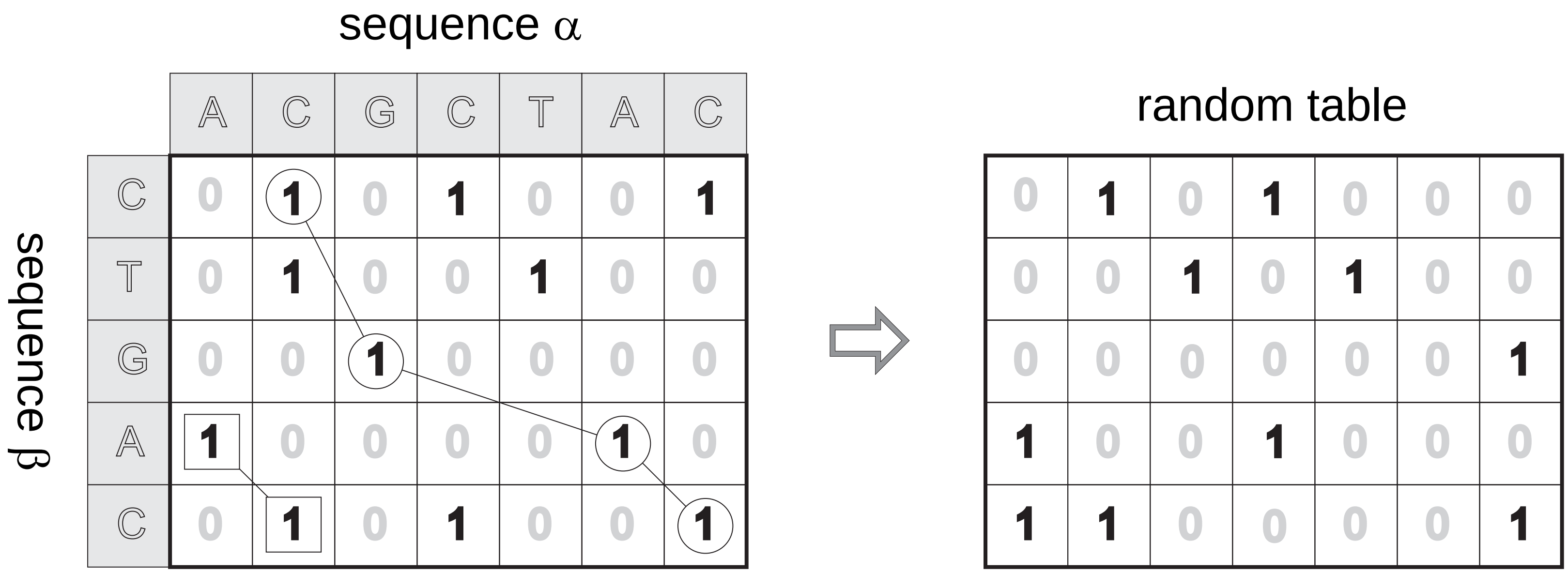
The variables $\eta_{i,j}$ **are not independent!**

Consider two words $\alpha = AB$ and $\beta = AA$

$$\left.\begin{array}{c}\eta_{1,1} = \eta_{1,2} = 1 \\ \eta_{2,1} = 0\end{array}\right\} \Rightarrow \eta_{2,2} = 0$$

Thus, the variables $\eta_{1,1}, \eta_{1,2}, \eta_{2,1}, \eta_{2,2}$ **are correlated.**

In the simplest (however still nontrivial) variant of the model, known as “Bernoulli Model”, it is supposed that all $\eta_{i,j}$ are independent uncorrelated **random variables** with the probability distribution $p=1/c$, where c is the number of letters in the alphabet.



	A	C	G	C	T	A	C
C	0	1	0	1	0	0	1
T	0	1	0	0	1	0	0
G	0	0	1	0	0	0	0
A	1	0	0	0	0	1	0
C	0	1	0	1	0	0	1

0	1	0	1	0	0	0
0	0	1	0	1	0	0
0	0	0	0	0	0	1
1	0	0	1	0	0	0
1	1	0	0	0	0	1

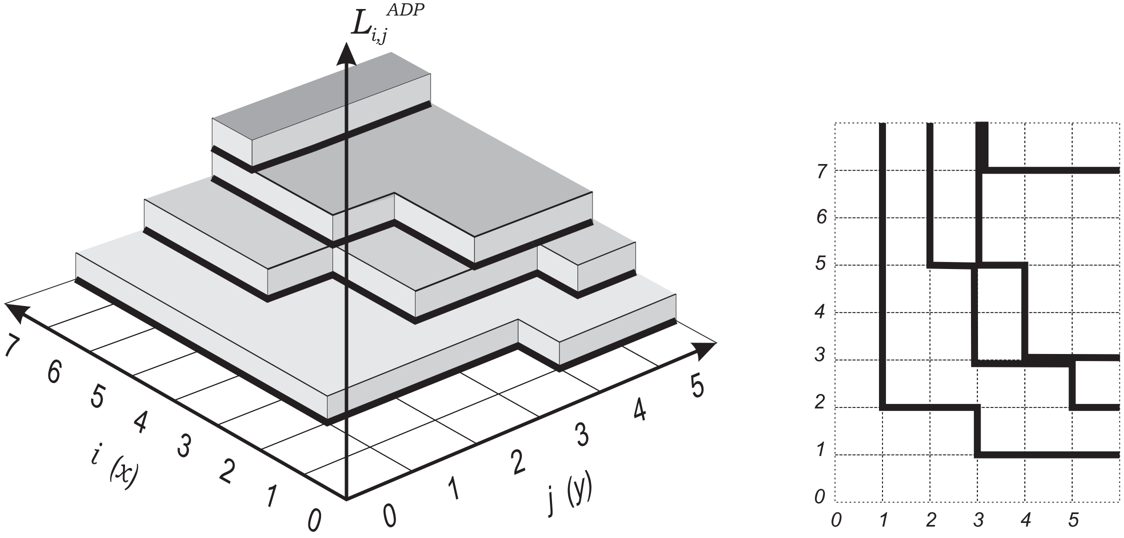
Connection of Bernoulli Matching with Directed Percolation in (2+1)-dimensional space

(2+1)-Anisotropic Directed Percolation (ADP) on a cubic lattice: the bonds along the axes x, y, z are occupied with the probabilities p_x, p_y, p_z.

Let us set $p_x = p_y = 1$ and $p_z = p$. If some point (x, y, z) belongs to the percolation cluster, then the point (x', y', z) , where $x' \geq x, y' \geq y$ also belongs to the cluster. **The cluster is compact** and is characterized by the height $L^{ADP}(x, y)$:

$$L^{ADP}(x, y) = \max\left[L^{ADP}(x-1, y), L^{ADP}(x, y-1) \right] + \xi_{x,y}$$

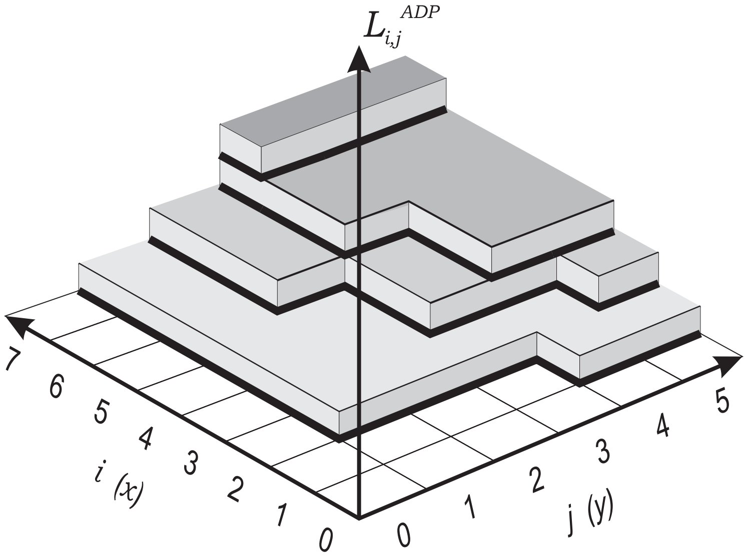
where $\text{Prob}\left(\xi_{x,y} = k\right) = (1-p)\,p^k \qquad (k = 0, 1, 2, \ldots)$



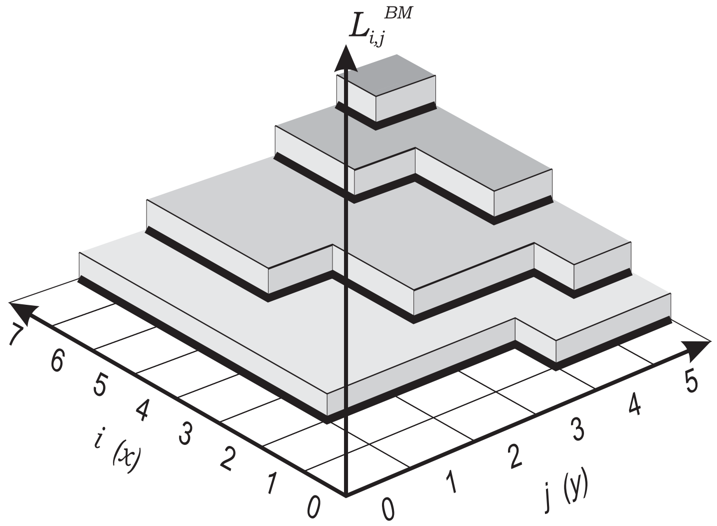
Compare the clusters in the Anisotropic Directed Percolation problem **(ADP)** and in the Bernoulli Matching **(BM)**

$$L^{ADP}(x,y)=\max\left[L^{ADP}(x-1,y),L^{ADP}(x,y-1)\right]+\xi_{x,y} \qquad L_{i,j}=\max\left[L_{i-1,j},L_{i,j-1},L_{i-1,j-1}+\eta_{i,j}\right]$$

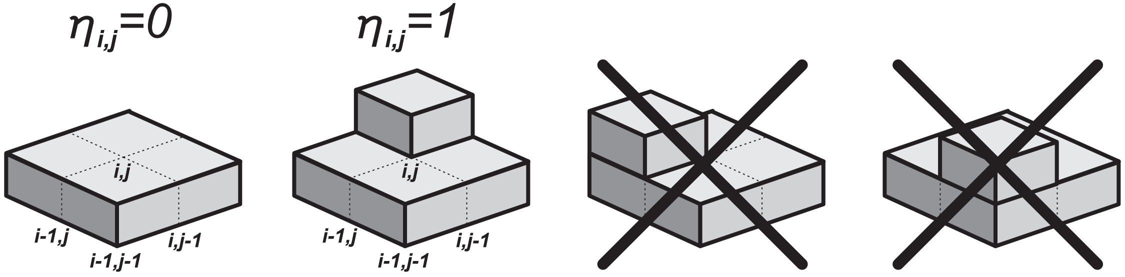
Anisotropic percolation



Bernoulli Matching

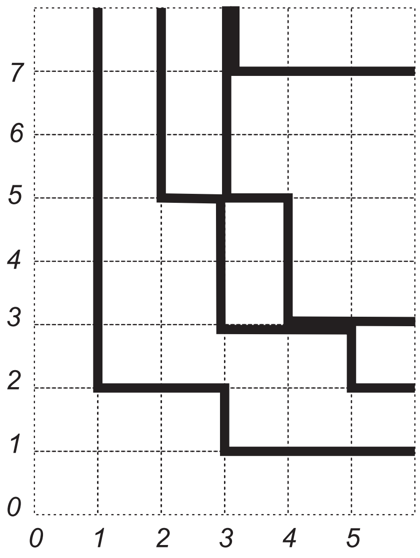


Bernoulli Matching



“World lines” representation of ADP and BM. These models are connected by a **nonlinear transform**

Anisotropic percolation

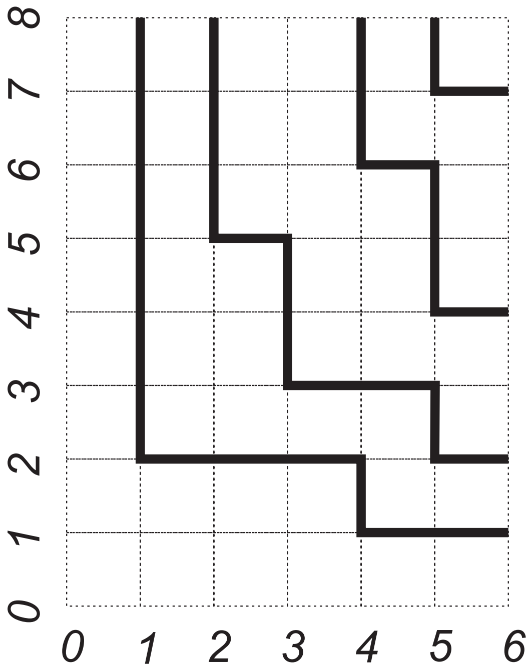


$$x_{BM} = x_{ADP} + L^{ADP}(x_{ADP}, y_{ADP})$$

$$y_{BM} = y_{ADP} + L^{ADP}(x_{ADP}, y_{ADP})$$

$$\Rightarrow$$

Bernoulli Matching



Formal connection between the cluster heights in ADP and BM models

$$L^{BM}(\zeta, \lambda) = L^{ADP}\left(\zeta - L^{BM}(\zeta, \lambda), \lambda - L^{BM}(\zeta, \lambda)\right)$$

Differential relations between the heights

If $\zeta = x + L^{ADP}(x,y); \quad \lambda = y + L^{ADP}(x,y)$, then for the derivatives (increments of heights) we get the following relations

$$\partial_x L^{ADP} = \frac{\partial_\zeta L^{BM}}{1 - \partial_\zeta L^{BM} - \partial_\lambda L^{BM}}; \quad \partial_y L^{ADP} = \frac{\partial_\lambda L^{BM}}{1 - \partial_\zeta L^{BM} - \partial_\lambda L^{BM}}$$

and vice versa

$$\partial_\zeta L^{BM} = \frac{\partial_x L^{ADP}}{1 + \partial_x L^{ADP} + \partial_y L^{ADP}}; \quad \partial_\lambda L^{BM} = \frac{\partial_y L^{ADP}}{1 + \partial_x L^{ADP} + \partial_y L^{ADP}}$$

These expressions are invariant with respect to the following transform

$$\zeta \to -x; \quad \lambda \to -y; \quad L^{BM} \to L^{ADP}$$

The height $L^{ADP}(x,y)$ in (2+1)-dimensional model of anisotropic directed percolation

Recursive relation

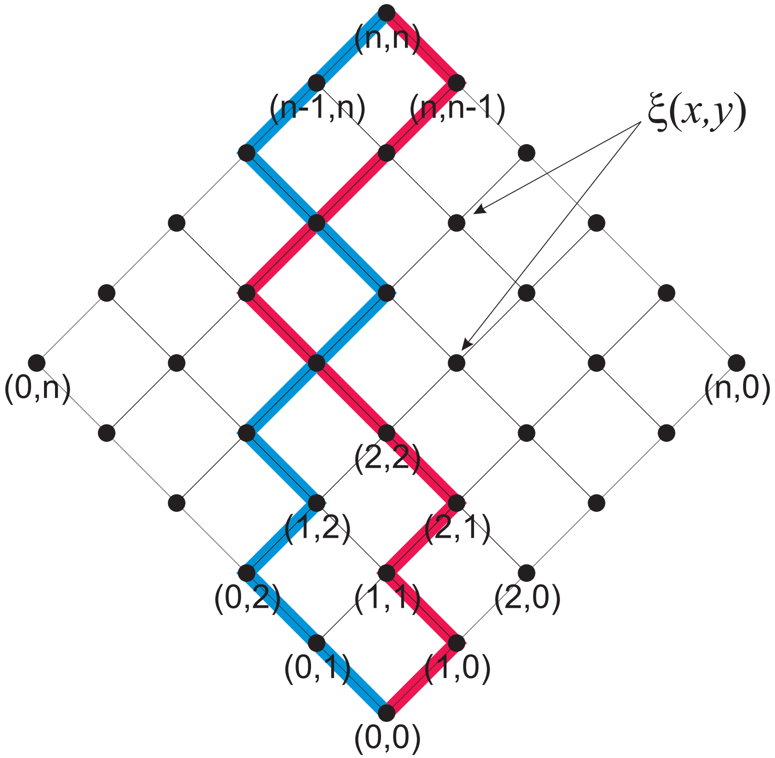
$$L^{ADP}(x,y) = \max\left[L^{ADP}(x-1,y), L^{ADP}(x,y+1)\right] + \xi_{x,y}$$

determines the **ground state energy** of directed polymer in a random environment with the Poisson distribution $\text{Prob}\left(\xi_{x,y}=k\right)=(1-p)\,p^{k}$

Statistical sum of the directed polymer in the random potential $\xi_{x,y}$ is

$$Z = \lim_{\beta\to\infty} \sum_{\text{over all paths}} \exp\left(-\beta \sum_{\text{along a path}} \xi(x,y) \right)$$

The asymptotic distribution of $L^{ADP}(x,y)$ is known (K. Johansson)



Asymptotic distribution of the **ground state energy** of directed polymer in a random Poissonian field

⇑

Asymptotic distribution of the **longest increasing subsequence** in a a random sequence of integers (“Ulam problem”)

⇑

Asymptotic distribution of the **first line** of the Young diagram over the Plancherel mesure
(distribution of the **largest eigenvalue** of random matrices from Gaussian ensemble)

$$L^{ADP}(x,y) \to \frac{2\sqrt{pxy}+p(x+y)}{1-p} + \frac{\left(pxy\right)^{1/6}}{1-p}\left[(1+p)+\sqrt{\frac{p}{xy}}(x+y)\right]^{2/3}\chi$$

where χ is the random variable distributed with the Tracy-Widom law.

Changing the variables

$$\zeta \to -x; \quad \lambda \to -y; \quad L^{BM} \to L^{ADP}$$

we get for Bernoulli Matching model

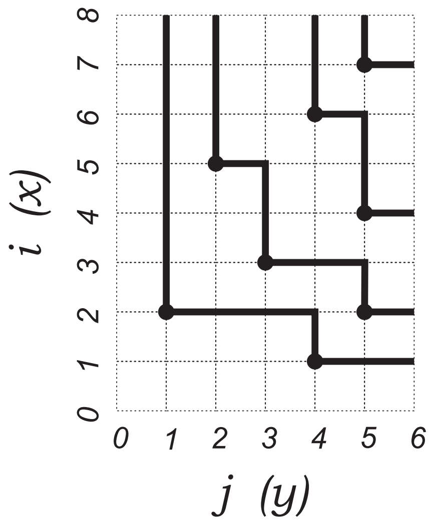
$$L^{BM}(x,y) \to \frac{2\sqrt{pxy} - p(x+y)}{1-p} + \frac{(pxy)^{1/6}}{1-p}\left[(1+p) - \sqrt{\frac{p}{xy}}(x+y)\right]^{2/3} \chi$$

If $x = y = N$, then the final expression has the form:

$$\begin{cases} \left\langle L^{BM} \right\rangle \approx \dfrac{2}{\sqrt{c}+1} N + \langle \chi \rangle \dfrac{c^{1/6}\left(\sqrt{c}-1\right)^{1/3}}{\sqrt{c}+1} N^{1/3} \\ \operatorname{Var} L^{BM} \approx \left(\left\langle \chi^2 \right\rangle - \langle \chi \rangle^2\right) \left(\dfrac{c^{1/6}\left(\sqrt{c}-1\right)^{1/3}}{\sqrt{c}+1}\right)^2 N^{2/3} \end{cases}$$

$$\langle \chi \rangle = -1.7711\ldots\,; \quad \left\langle \chi^2 \right\rangle - \langle \chi \rangle^2 = 0.8132\ldots$$

Computation of averaged $L_{i,j}$ from simple consideration and from Bethe ansatz



The statistical weight of the configuration C is

$$W(C) = p^{N_o} q^{N_h}$$

where N_o, N_h is the number of occupied sites ("particles") and holes ("empty sites").

Let $P(s\,|\,m)$ is the probability to jump on s steps under the condition that the next particle is located on the same line to the right on the distance in m steps

$$P(s\,|\,m) = p^{1-\delta_{s,0}} q^{m-1-s} \qquad \left(\sum_{s=0}^{m=1} P(s\,|\,m) = 1 \right)$$

The mean value $d(m) \equiv \langle s \rangle = \sum_{s=0}^{m=1} sP(s\,|\,m)$ over the distribution $P(s\,|\,m)$ is

$$d(m) = m - \frac{1-q^m}{p}$$

Let ρ be the density of the particles in the system (ρ = const). In the stationary regime the probability $Q(m)$ to find a distance between the neighboring particles equal to *m*, *is*

$$Q(m) = \rho(1-\rho)^{m-1}$$

Averaging now $d(m)$ over the distribution $Q(m)$, we get an averaged length of a jump in a system of particles with the concentration ρ:

$$\boxed{\langle d \rangle = \sum_{\mathrm{m=1}}^{\infty} d(m) Q(s\,|\,m) = \frac{p(1-\rho)}{\rho(p+q\rho)}}$$

Define the height increment:

$$\partial_y L = \begin{cases} 0 & \text{if we do not cross the line} \\ 1 & \text{if we cross the line} \end{cases}$$

The average increment of the height in horizontal direction per one step in vertical direction is:

$$\begin{cases} \dfrac{\partial_x L}{\partial_y L} = \langle d \rangle \\ \partial_y L = \rho \end{cases} \Rightarrow \boxed{p(1-\partial_y L - \partial_x L) = q\,\partial_x L\,\partial_y L}$$

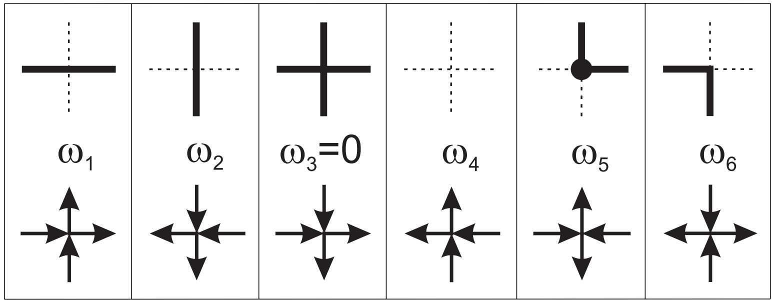
The solution of this differential equation reads:

$$L(x,y) = \frac{2\sqrt{pxy} - p(x+y)}{1-p}$$

At $x = y = N$ we get an answer for the average length of a subword in the alphabet of c letters ($p = 1/c$):

$$L(x,y) = \frac{2}{\sqrt{c}+1} N$$

The same result we can obtain from the exact relations of Bethe ansatz



The Boltzmann weights

$$\omega_1 = e^{\mu}; \quad \omega_2 = 1; \quad \omega_3 = 0;$$

$$\omega_4 = q = 1 - p; \quad \omega_5 \omega_6 = pe^{\mu}$$

The statistical sum of the grand canonical ensemble of 5-vertex model is as follows:

$$Z = \sum_{\text{configurations}} \omega_1^{N_h} \omega_2^{N_v} \omega_4^{N_e} \left(\omega_5 \omega_6 \right)^{N_c}$$

where N_h, N_v, N_e, N_c are the numbers of horizontal and vertical bonds, empty sites and corners (for each configuration of lines in the system).

Knowing the statistical sum, one can compute the average flux

$$\bar{\Phi} \equiv \left\langle N_h \right\rangle = \frac{\sum N_h (1-p)^{N_e} (e^{\mu})^{N_h} (pe^{\mu})^{N_c}}{\sum (1-p)^{N_e} (e^{\mu})^{N_h} (pe^{\mu})^{N_c}} = \frac{\partial}{\partial \mu} \ln Z(p, \mu) \Big|_{\mu=0}$$

The Bethe equations read:

$$\begin{cases} Z = \left(\Lambda_n\right)^N \\ \Lambda_n = \omega_2^n \omega_4^{N-n} \prod_{j=1}^{N-n} \left(1 + \dfrac{\omega_5\omega_6}{\omega_2\omega_4} z_j\right), \qquad \left(j = 1, \ldots, N-n\right) \\ z_j = (-1)^{N-n-1} \prod_{i=1}^{N-n} \dfrac{1-\Delta z_j}{1-\Delta z_i}, \qquad \Delta = \dfrac{\omega_1\omega_2 - \omega_5\omega_6}{\omega_2\omega_4} = e^{\mu} \end{cases}$$

The solution of Bethe equations leads to the following answer (where $\rho = \dfrac{n}{N}$):

$$\frac{1}{N} \ln Z(p,\mu) = \frac{p(1-\rho)}{p+q\rho} \mu N + \frac{\sqrt{\pi}}{4} \frac{p}{p+q\rho} \frac{(1-\rho)^{3/2}}{\rho^{1/2}} \mu^2 N^{3/2} + \ldots$$

what gives for the average flux already known expression

$$\boxed{\bar{\Phi} \equiv \left\langle N_h \right\rangle = \frac{p(1-\rho)}{\rho(p+q\rho)}}$$