



Ultrametric models in theory of symbolic sequences.

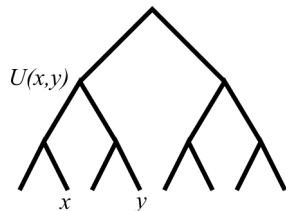
Vladimir A.I. Osipov

October 25, 2017

p -adic diffusion equation:

$$\frac{\partial f(x,t)}{\partial t} = D_x^\beta f(x,t)$$

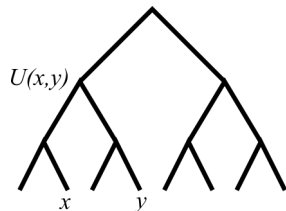
where $x \in \mathbb{Q}_p$, $t \in \mathbb{R}$, $\beta \sim \frac{1}{T}$. Offers an accurate and universal description of the **protein conformation dynamics**. The description takes into account the symmetry properties (hierarchical self-similarity) of the state space only.



p -adic diffusion equation:

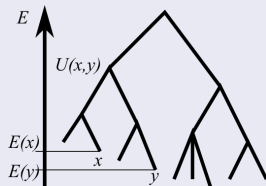
$$\frac{\partial f(x,t)}{\partial t} = D_x^\beta f(x,t)$$

where $x \in \mathbb{Q}_p$, $t \in \mathbb{R}$, $\beta \sim \frac{1}{T}$. Offers an accurate and universal description of the **protein conformation dynamics**. The description takes into account the symmetry properties (hierarchical self-similarity) of the state space only.

**The ultrametric diffusion equation with a drift term:**

$$\frac{\partial f(x,t)}{\partial t} = \int_x d\mu(y) \frac{e^{-\beta U(|x,y|)}}{|x,y|} \left(e^{\beta E(y)} f(y,t) - e^{\beta E(x)} f(x,t) \right)$$

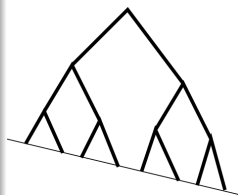
It describes diffusion on an arbitrary hierarchical energy landscape. The additional terms generate a drift in the space.



Block-rectangular hierarchical matrix:

$$\left(\begin{array}{cc|c|cc} q_0 & q_1 & & q_0 & q_1 \\ \hline q_0 & q_1 & & q_0 & q_1 \\ q_1 & q_0 & q_2 & q_1 & q_0 \\ \hline q_1 & q_0 & & q_1 & q_0 \\ \hline q_2 & & q_0 & q_1 & \\ & & \hline & & q_0 & q_1 \\ & & q_1 & q_0 & \\ & & \hline & & q_1 & q_0 \end{array} \right) .$$

Here q_2 is a 4×2 matrix. The model describes ultrametric diffusion with a constant drift.



Rigorous definition

Tensor product representation of Parisi matrix.

Hilbert 2^r dimensional space: $\mathcal{H}_{2^r} = h \otimes h \otimes \cdots \otimes h$, let $\omega \in \mathcal{H}_{2^r}$.

$$\hat{Q}\omega = \sum_{\gamma=0}^r a_{\gamma} \hat{S}_{\gamma} \omega, \quad \hat{S}_{\gamma} \omega = \underbrace{\omega_1 \otimes \cdots \otimes \omega_{r-\gamma}}_{r-\gamma} \otimes \underbrace{\hat{S}\omega_{r-\gamma+1} \otimes \cdots \otimes \hat{S}\omega_r}_{\gamma}$$

where a_{γ} 's are arbitrary numbers and \hat{S} is a projection operator,
 $\hat{S}|1\rangle = |1\rangle$, $\hat{S}|0\rangle = 0|0\rangle$.

The operator is a Parisi matrix in the basis composed of $|0\rangle$ and $|1\rangle$
 and has eigenvalues $\lambda^{(\mu)} = \sum_{\gamma=0}^{\mu} a_{\gamma}$ with $\text{mult}(\lambda^{(\mu)}) = 2^{r-\mu-1}$.

Rigorous definition

Translation operator \hat{T} . The acts on vectors from \mathcal{H} as follows

$$\hat{T}\omega_1 \otimes \cdots \otimes \omega_{r-1} \otimes \omega_r = \omega_r \otimes \omega_1 \otimes \cdots \otimes \omega_{r-1}, \quad \omega_j \in h.$$

$$\hat{T} = \begin{pmatrix} |1\rangle & & & & |0\rangle & & & & \\ & |1\rangle & & & & |0\rangle & & & \\ & & \ddots & & & & \ddots & & \\ & & & \ddots & & & & \ddots & \\ & & & & \ddots & & & & \ddots \\ & & & & & |1\rangle & & & \\ & & & & & & & & |0\rangle \end{pmatrix},$$

Rigorous definition

Translation operator \hat{T} . The acts on vectors from \mathcal{H} as follows

$$\hat{T}\omega_1 \otimes \cdots \otimes \omega_{r-1} \otimes \omega_r = \omega_r \otimes \omega_1 \otimes \cdots \otimes \omega_{r-1}, \quad \omega_j \in h.$$

$$\hat{T} = \begin{pmatrix} |1\rangle & & & & |0\rangle & & & & \\ & |1\rangle & & & & |0\rangle & & & \\ & & \ddots & & & & \ddots & & \\ & & & \ddots & & & & \ddots & \\ & & & & |1\rangle & & & & |0\rangle \end{pmatrix},$$

New family of operators is introduced as the products

$$\hat{Q} = \hat{T}\hat{Q},$$

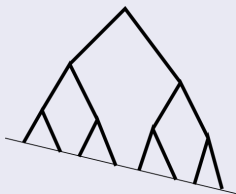
Thus the evolution generated by \hat{Q} is a composition of ultrametric diffusion \hat{Q} and deterministic process induced by \hat{T} .

New family of operators

$$\hat{Q} = \hat{T}\hat{Q},$$

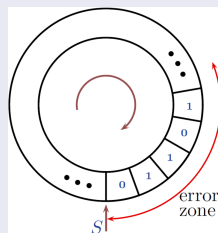
Thus the evolution generated by \hat{Q} is a composition of ultrametric diffusion \hat{Q} and deterministic process induced by \hat{T} .

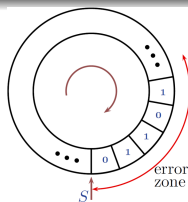
Ultrametric diffusion with a constant drift



Initial stage of the process.

Errors generation model in information sequences





- The information carrier ("disk") consists of r cells. Each of them contains a digit $\{0, 1\}$. At the discrete moments of time $t = 0, 1, \dots$ disc rotates one step clockwise, at $t = \ell \leq r$ the pointer S points to the ℓ th cell.
- The noise affects the cells placed "rightwards" to the point S . The probability to affect γ consecutive cells is given by a_γ ($\sum_{\gamma=0}^r a_\gamma = 1$). The content of each affected cell is replaced by any symbol from $\{0, 1\}$ with equal probability.
- Each state of the disc can be described by a vector $\mathbf{w} \in \mathcal{H}_{2^r}$:

$$\mathbf{w}(t) = \sum_J w_J(t) |e_J\rangle, \quad |e_J\rangle = \bigotimes_{k=1}^r |j_k\rangle, \quad j_k \in \{0, 1\},$$

where $w_J(t)$ are probabilities to find the disc having informational content encoded by the sequence of symbols $J = [j_1 j_2 \dots j_r]$.

- The result of ℓ steps time evolution is described by the operator

$$\prod_{i=1}^{\ell} \left[\hat{T} \sum_{\gamma=0}^r a_\gamma \hat{S}_\gamma \right] = (\hat{T} \hat{Q})^\ell = \hat{Q}^\ell$$

Spectral problem.

The spectral problem for the matrix \hat{Q}^r

$$\Lambda^{(m;\nu)} = \begin{cases} (\lambda^{(m)})^r, & m = 0, r; \\ (\lambda^{(0)})^{r-m} \prod_{i=1}^{\ell} (\lambda^{(i)} \lambda^{(i-1)} \dots \lambda^{(1)})^{\nu_i}, & 0 < m < r. \end{cases}$$

The corresponding multiplicities are given by

$$\text{mult} \left(\Lambda^{(m;\nu)} \right) = \begin{cases} 1, & m = 0, r; \\ \frac{r(r-m-1)!}{(r-m-\sum_i \nu_i)! \prod_i \nu_i!}, & 0 < m < r, \end{cases}$$

where $\lambda^{(\mu)} = \sum_{\gamma=0}^{\mu} a_{\gamma}$ and ν_i is the constrained partition of m : $\sum_{i=1}^{\ell} i\nu_i = m$, such that $\sum_{i=1}^{\ell} \nu_i \leq r - m$.

Estimation of losses.

The moment generation function

An auxiliary operator \hat{E}

$$\hat{E} = (\mathbf{1} + e^\alpha \mathbf{P}) \otimes (\mathbf{1} + e^\alpha \mathbf{P}) \otimes \cdots \otimes (\mathbf{1} + e^\alpha \mathbf{P});$$

$$\mathbf{1} = |0\rangle\langle 0| + |1\rangle\langle 1|, \quad \mathbf{P} = |0\rangle\langle 1| + |1\rangle\langle 0|,$$

so that $\langle J | \hat{E} | J' \rangle = e^{k\alpha}$, with k being the number of different symbols (compared pairwise) in the sequences J and J' .

The moment generation function

$$R_\ell(\alpha) = \langle J_0 | \hat{T}^{-\ell} \hat{E} \hat{Q}^\ell | J_0 \rangle .$$

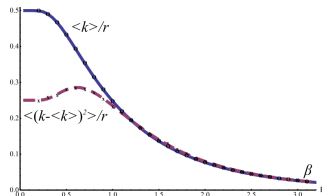
The mean value of the changes accumulated in the disk after ℓ rotations and its square displacement are

$$\langle k \rangle_{t=\ell} = \partial_\alpha R_\ell(\alpha)|_{\alpha=0};$$

$$\langle (k - \langle k \rangle)^2 \rangle_{t=\ell} = \partial_\alpha^2 R_\ell(\alpha)|_{\alpha=0} - \langle k \rangle_{t=\ell}^2 .$$

Case of Boltzmann noise.

The probabilities a_γ are proportional to $e^{-\beta\gamma}$ with parameter $\beta \sim 1/T$ (inverse temperature).

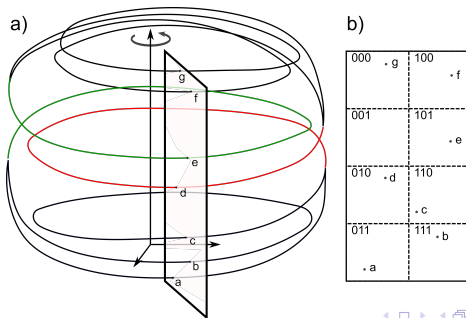


- $\beta \ll 1$: At high temperatures $\langle k \rangle \sim \frac{r}{2}$ and $\text{var}(k) \sim \frac{r}{4}$. Almost every cell are affected. The probability that exactly k cells change their content is $2^{-r} \binom{r}{k}$ (Binomial distribution).
- $\beta \gg 1$: The randomising events caused by the noise are rare and mostly affect one cell only. One observes the equality $\langle k \rangle \simeq \text{var}(k)$ (Poisson process).
- $\beta \sim 1$: The variance reaches its maximal value.

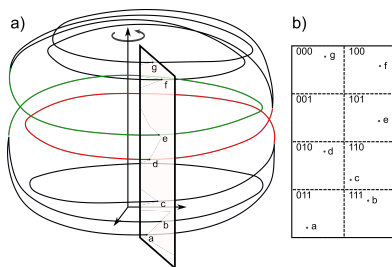
Ultrametric models in symbolic sequences theory

Symbolic sequence is a fundamental concept

- In bioinformatics, information theory, discrete Markov chains. There is a direct mapping of physical object structure on the set of symbolic sequences.
- Dynamical system. The real trajectory can be restored from symbolic dynamics.



Symbolic dynamics – stroboscopic sampling of the multidimensional trajectory



1. Poincaré section surface. In Hamiltonian dynamics the surface is orthogonal to the dynamical flow at each point of phase-space.
 2. Linearisation allows to define the set of feasible positions at the next crossing.
 3. All intersections falling within the same region are designated by a certain symbol.
- For infinite or cyclic symbolic sequence the real trajectory can be restored uniquely.
 - There are chaotic systems with finite Markov alphabet (Backer's Map).

Periodic orbits in Quantum chaos

- One interested in statistics (correlations) of energy levels of a quantum system which is chaotic in a classical limit.
- Gutzwiller trace formula ($\hbar \rightarrow 0$)

$$\rho(E) = \sum_n \delta(E - \varepsilon_n) \simeq \underbrace{W(E)}_{\text{smooth part}} + \text{Re} \sum_{\gamma \in P.O.} \mathcal{A}_\gamma \exp\left(\frac{i}{\hbar} S_\gamma(E)\right)$$

- Form factor **correlations of** $\varepsilon_n \Leftrightarrow$ **correlations of** S_γ

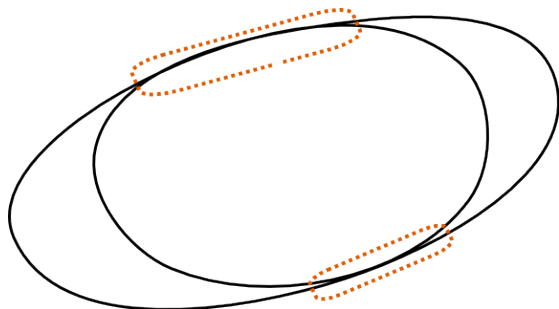
$$\mathcal{K}(\tau) = \int d\varepsilon e^{-i\varepsilon\tau} \langle \rho(E+\varepsilon)\rho(E-\varepsilon) \rangle \simeq \sum_{\gamma\gamma'} \mathcal{A}_\gamma \mathcal{A}_{\gamma'}^* \exp\left(\frac{i}{\hbar} \Delta S_{\gamma\gamma'}\right)$$

- Classification of trajectories with respect to their contribution.

$$\mathcal{K}(\tau) = \int d\varepsilon e^{-i\varepsilon\tau} \langle \rho(E+\varepsilon)\rho(E-\varepsilon) \rangle \simeq \sum_{\gamma\gamma'} \mathcal{A}_\gamma \mathcal{A}_{\gamma'}^* \exp\left(\frac{i}{\hbar} \Delta S_{\gamma\gamma'}\right)$$

Classification of trajectories on their contribution at $\hbar \rightarrow 0$:

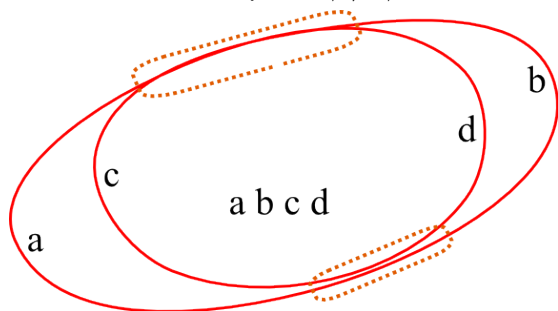
1. diagonal approximation by Berry, $\gamma = \gamma'$;
2. Sieber-Richter pairs, $\gamma \neq \gamma' \ni$ encounter;



$$\mathcal{K}(\tau) = \int d\varepsilon e^{-i\varepsilon\tau} \langle \rho(E+\varepsilon)\rho(E-\varepsilon) \rangle \simeq \sum_{\gamma\gamma'} \mathcal{A}_\gamma \mathcal{A}_{\gamma'}^* \exp\left(\frac{i}{\hbar} \Delta S_{\gamma\gamma'}\right)$$

Classification of trajectories on their contribution:

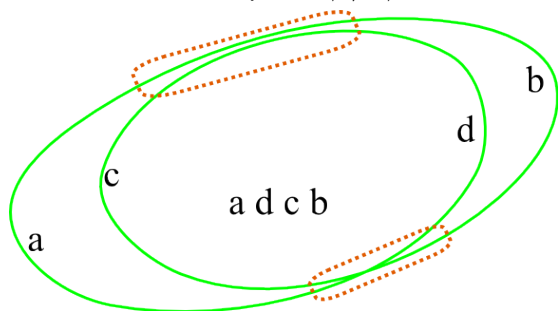
1. diagonal approximation by Berry, $\gamma = \gamma'$;
2. Sieber-Richter pairs, $\gamma \neq \gamma' \ni$ encounter;



$$\mathcal{K}(\tau) = \int d\varepsilon e^{-i\varepsilon\tau} \langle \rho(E+\varepsilon)\rho(E-\varepsilon) \rangle \simeq \sum_{\gamma\gamma'} \mathcal{A}_\gamma \mathcal{A}_{\gamma'}^* \exp\left(\frac{i}{\hbar} \Delta S_{\gamma\gamma'}\right)$$

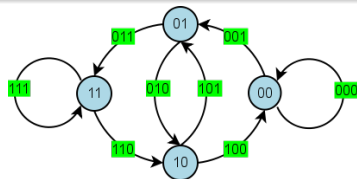
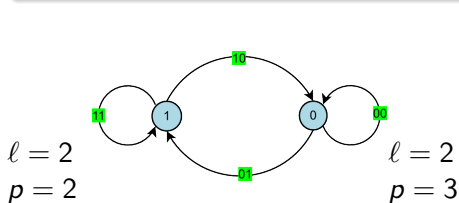
Classification of trajectories on their contribution:

1. diagonal approximation by Berry, $\gamma = \gamma'$;
2. Sieber-Richter pairs, $\gamma \neq \gamma' \ni$ encounter;

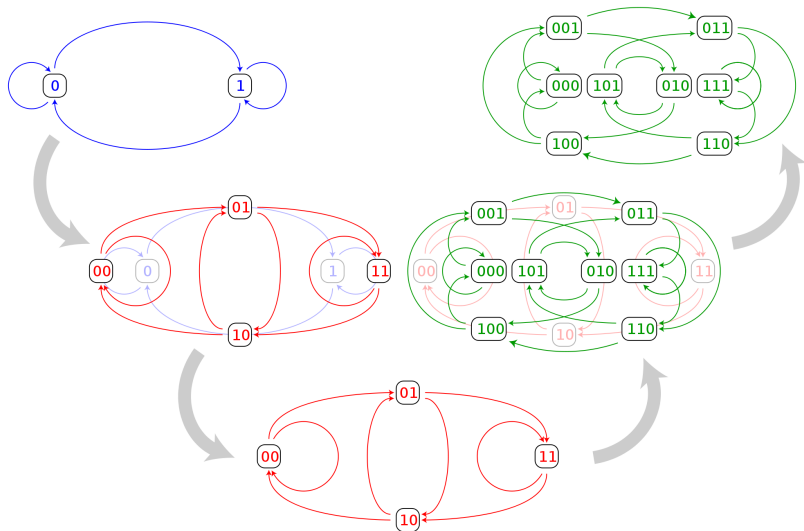


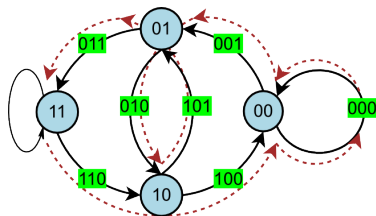
De Bruijn graph of order p on ℓ symbols

- directed graph.
- vertices are strings of $p - 1$ symbols from ℓ -letters alphabet.
- edges are p strings: $\{a_1, a_2, \dots, a_{p-1}\} \rightarrow \{a_2, \dots, a_{p-1}, a_p\}$

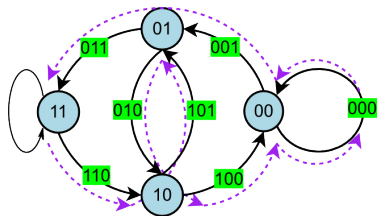


De Bruijn graph generations

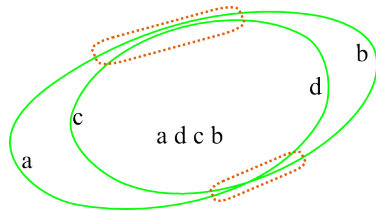
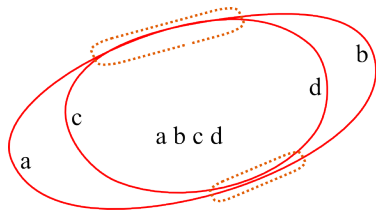




$$\gamma = [11000101]$$



$$\gamma' = [11010001]$$



p -close sequences

p -closeness ($\gamma \stackrel{p}{\sim} \beta$)

Two sequences, γ and β , are p -close if any substring of the length $p - 1$ appears the same number of times in both γ and β .

p -close sequences

p -closeness ($\gamma \stackrel{p}{\sim} \beta$)

Two sequences, γ and β , are p -close if any substring of the length $p - 1$ appears the same number of times in both γ and β .

Properties of $\stackrel{p}{\sim}$

- If $\gamma \stackrel{p}{\sim} \alpha$ and $\gamma \stackrel{p}{\sim} \beta$, then $\alpha \stackrel{p}{\sim} \beta$ (transitivity);
- If $\gamma \stackrel{p+1}{\sim} \alpha$ then $\gamma \stackrel{p}{\sim} \alpha$ (nesting).

p -close sequences

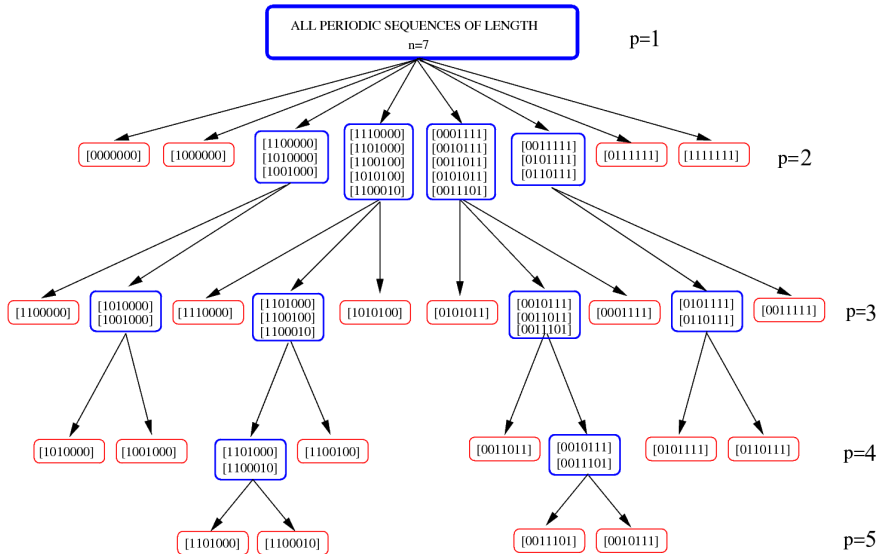
p -closeness ($\gamma \stackrel{p}{\sim} \beta$)

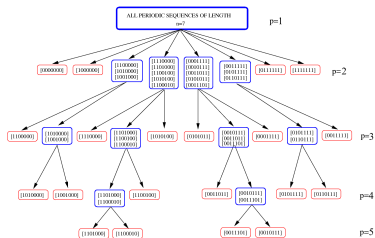
Two sequences, γ and β , are p -close if any substring of the length $p - 1$ appears the same number of times in both γ and β .

Properties of $\stackrel{p}{\sim}$

- If $\gamma \stackrel{p}{\sim} \alpha$ and $\gamma \stackrel{p}{\sim} \beta$, then $\alpha \stackrel{p}{\sim} \beta$ (transitivity);
- If $\gamma \stackrel{p+1}{\sim} \alpha$ then $\gamma \stackrel{p}{\sim} \alpha$ (nesting).

All symbolic sequences of a given length can be distributed over hierarchically nested clusters.



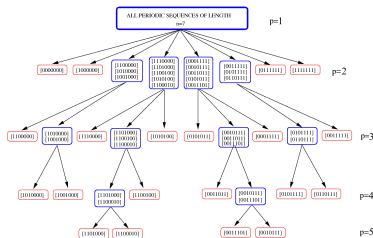


Define ultrametric distance:

$$d(\alpha, \beta) = e^{-p_{\max}} \quad p_{\max} = \max(p)_{\alpha \sim \beta}$$

$$d(\alpha, \beta) \leq d(\alpha, \gamma) + d(\beta, \gamma);$$

$$d(\alpha, \beta) \leq \max \left(d(\alpha, \gamma), d(\beta, \gamma) \right).$$



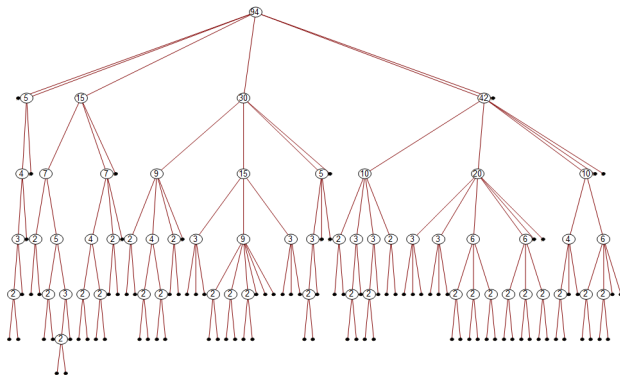
Define ultrametric distance:

$$d(\alpha, \beta) = e^{-p_{\max}} \quad p_{\max} = \max(p)_{\alpha \sim \beta}$$

$$d(\alpha, \beta) \leq d(\alpha, \gamma) + d(\beta, \gamma);$$

$$d(\alpha, \beta) \leq \max \left(d(\alpha, \gamma), d(\beta, \gamma) \right).$$

What are statistics of clusters?

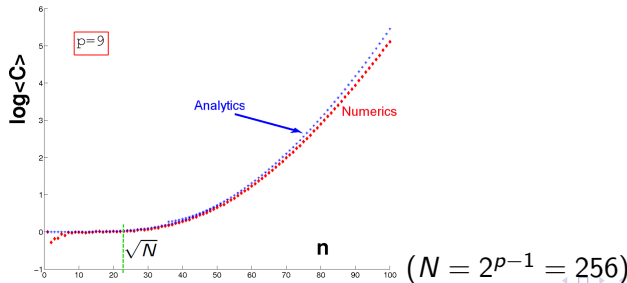


Clustering of cyclic binary sequences of the length $n = 11$ (only half of the tree, 94 sequences out of 188, is shown). The end-points represent the sequences, the numbers in the circles give the total number of sequences in the corresponding cluster.

The maximal branching level of the tree is $p_{max} = \lceil \frac{n-3}{2} \rceil + 1$.

Some results regarding the cluster sizes

- The total number of clusters in the regime of finite p and $n \gg 1$ grows as $\frac{1}{4}n^{2^{p-1}}(1 + O(1/n))$
 - The asymptotic for the average sizes of clusters, $\langle \|\mathcal{C}\| \rangle$.
 - For $n \lesssim \sqrt{2^p}$ the average cluster size is one, $\langle \|\mathcal{C}\| \rangle \approx 1$, and the number of clusters is almost equal to the total number of cyclic sequences.
 - For $2^{p/2} \ll n \ll 2^p$ one can find that $\log \langle \|\mathcal{C}\| \rangle \sim n^2$.
 - For $n \gg 2^p$ the average size is $\langle \|\mathcal{C}\| \rangle = \frac{2^n}{n} \left(\frac{2^{p-1}}{n\pi} \right)^{2^{p-2}} (1 + O(1/n))$.
- Further growth of n gives $\log \langle \|\mathcal{C}\| \rangle \sim n$.

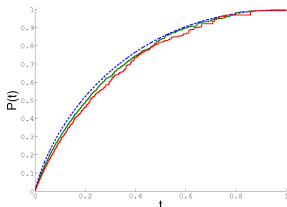


Some results regarding the cluster distributions

- The size of the largest cluster, $\|C_{\max}\|$ in the limit of long sequences $n \gg 1$ and finite p behaves like $\|C_{\max}\| = \left(\frac{2^n}{n}\right) \left(\frac{2^p}{\pi n}\right)^{2^{p-2}} (1 + O(n^{-1}))$.
- The probability density that a randomly chosen cyclic sequence belongs to a cluster of the size less then $t \|C_{\max}\|$, $t \in [0, 1]$, it is ($n \gg 1$)

$$\rho(t) = \frac{(-\log t)^{2^{p-2}-1}}{(2^{p-2} - 1)!}.$$




- Probability to find k randomly chosen sequences of the same length to be belonging to the same cluster: $\left(\frac{1}{k}\right)^{2^{p-2}} \left(\frac{2^p}{\pi n}\right)^{(k-1)2^{p-2}} (1 + O(n^{-1}))$.



Exact distribution of cluster sizes at $n = 70$ (upper green) and $n = 47$ (lower red) is shown in comparison with the asymptotic expression $P(t) = t(1 - \log t)$ (dashed blue line) for the case $p = 3$.

- A model of operator of ultrametric diffusion with a constant drift has been proposed. The operator is effectively one acting on a set of symbolic sequences.
- A notion of p -closeness has been introduced. It generates an ultrametric distance on the set of cyclic symbolic sequences.
- – The *Hamming distance* between two given sequences, such that both have the same length and a fixed starting point, is defined as the total number of different symbols appearing in the identical positions. The Hamming distance does not take into account the surrounding symbols.
 - The *ultrametric distance* is insensitive to the absolute position of symbols within the sequence, while it takes into account the local surrounding. Thus even one different symbol in two sequences put them on the furthest distance from each other.
 - This prompts us the next natural step in the direction of classification of symbolic sequences. This should be combination of both approaches together.

Related publications

-  Gutkin B., Osipov V.A.I. “Spectral problem of block-rectangular hierarchical matrices” *Journal of Statistical Physics* **143** (2011) 72,
-  Gutkin B., Osipov V.A.I. “Clustering of periodic orbits in chaotic systems”, *Nonlinearity* **26** (2013) 177,
-  Osipov V.A.I. “Wavelet analysis on symbolic sequences and two-fold de Bruijn sequences”, *Journal of Statistical Physics* **164** (2016) 142