# Исследование свойств стохастических КС-грамматик вида «цепочки»

Выполнил: Мартынов И.М.
Научный руководитель: д.ф.-м.н., проф. Жильцова Л.П.

# Стохастическая КС-грамматика

$$G =< T, N, S, R >$$

$T$ и $N$ — конечные алфавиты терминальных и нетерминальных символов.

$$T = \{a, b, c, \ldots\}, \quad N = \{A_1, A_2, \ldots, A_n\}$$
$$T \cap N = \varnothing$$

$S$ — аксиома грамматики, $S \in N$.
$R$ — множество правил вывода. В стохастической грамматике каждому правилу приписывается вероятность.

$$R = \bigcup_{i=1}^{n} R_i \quad : \quad R_i = \{A_i \xrightarrow{p_{ij}} \beta_{ij}, \;\; j = \overline{1, n_i}\}$$

$$p_{ij} > 0, \quad \sum_{j=1}^{n_i} p_{ij} = 1$$

# Пример: язык Дика

$$T = \{a, b\}, \quad N = \{A\}$$

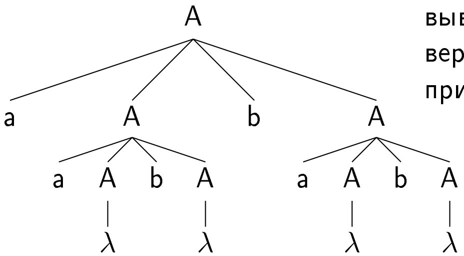
$$A \xrightarrow{1/4} aAbA$$

$$A \xrightarrow{3/4} \lambda$$

Вывод слова *aabbab* в грамматике:

$$A \to aAbA \to aaAbAbA \to aabAbA \to$$
$$\to aabbA \to aabbaAbA \to aabbabA \to aabbab$$

Дерево вывода:




Вероятность $p(d)$ дерева вывода $d$ — произведение вероятностей всех применённых правил.

$$\lim_{t\to\infty} \sum_{d \in D^{\leqslant t}} p(d) = 1.$$

# Классы нетерминалов, разложимость

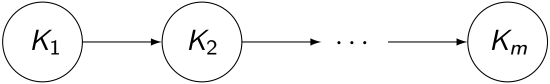
$A_i \to_* A_j$, если $A_i \to \beta_1 \to \beta_2 \to \ldots \to \beta_k \ni A_j$.

Множество нетерминалов разбивается на классы $K_1, K_2, \ldots, K_m$, такие что в любом классе $K_l$ для любых $A_i, A_j \in K_l$ выполняется $A_i \to_* A_j$ и $A_j \to A_i$.

$K_i \prec K_j$, если для каких-нибудь $A_l, A_k$ $A_l \in K_i$, $A_k \in K_j$, $A_l \to A_k$.

Будем рассматривать грамматики, имеющие вид «цепочки»:

$$K_l \prec K_h \Leftrightarrow h - l = 1$$



# Производящие функции, моменты

$$F_i(s_1, s_2, \ldots, s_n) = F_i(\mathbf{s}) = \sum p_{ij} s_1^{l_1} s_2^{l_2} \ldots s_n^{l_n}$$

$l_k = l_k(i,j)$ — число нетерминалов $A_k$ в правой части правила $A_i \xrightarrow{p_{ij}} \beta_{ij}$

$$F_i(s_1, s_2, \ldots, t) = \begin{cases} F_i(s_1, s_2, \ldots), t = 1 \\ F_i(\mathbf{F}(\mathbf{s}), t-1), t > 1 \end{cases},$$

где $\mathbf{F}(\mathbf{s}) = (F_1(s_1, \ldots, s_n), \ldots, F_n(s_1, \ldots, s_N))$.

Первые и вторые моменты:

$$\left.\frac{\partial F_i(s_1, \ldots s_n)}{\partial s_j}\right|_{\mathbf{s}=\mathbf{1}} = a_j^i \qquad \left.\frac{\partial^2 F_i(s_1, \ldots s_n)}{\partial s_k \partial s_j}\right|_{\mathbf{s}=\mathbf{1}} = b_{jk}^i$$

# Моменты

$A = (a^i_j)$ — матрица первых моментов

Для случая цепочки

$$A = \begin{pmatrix} A_{11} & A_{12} & 0 & \cdots & 0 & 0 \\ 0 & A_{22} & A_{23} & \cdots & 0 & 0 \\ 0 & 0 & A_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_{m-1,m-1} & A_{m-1,m} \\ 0 & 0 & 0 & \cdots & 0 & A_{mm} \end{pmatrix}$$

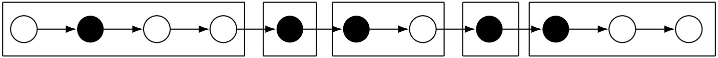
$r_i$ — перронов корень диагонального блока $A_{ii}$.

$r$ — перронов корень всей матрицы $A$.

$r = \max\{r_1, r_2, \ldots, r_m\}$.

Будем рассматривать случай $r = 1$ (критический случай).

# Асимптотика матрицы первых моментов



$$A^t = \begin{pmatrix} B_{11}^t & B_{12}^{(t)} & \cdots & B_{1w}^{(t)} \\ 0 & B_{22}^t & \cdots & B_{2w}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_{ww}^t \end{pmatrix}$$

$$B_{lh}^{(t)} = H_{lh} \cdot t^{s_{lh}-1} \cdot r^t \cdot (1 + o(1))$$

*Результат Л.П. Жильцовой*

В критическом случае,

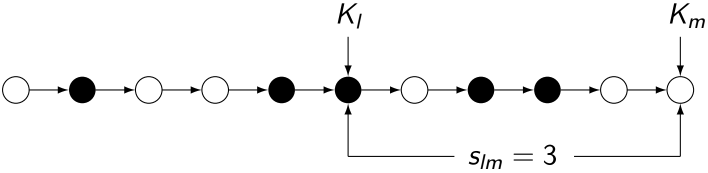
$$B_{lh}^{(t)} = H_{lh} \cdot t^{s_{lh}-1} \cdot (1 + o(1))$$

# Вероятности продолжения и деревьев фиксированной высоты

$Q_i(t)$ — вероятность того, что дерево вывода, построенное из $A_i$, продолжится на уровне $t$.
$P_i(t)$ — вероятность деревьев высоты $t$.

$$Q_i(t) = c_i t^{-\left(\frac{1}{2}\right)^{s_{lm}}} \cdot (1 + o(1))$$
$$P_i(t) = d_i t^{-1+\left(\frac{1}{2}\right)^{s_{lm}}} \cdot (1 + o(1)),$$

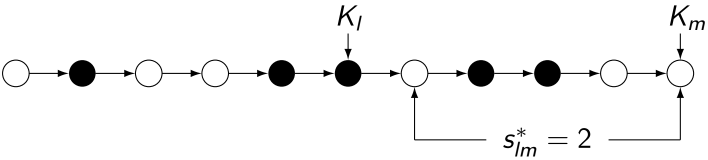
где $A_i \in K_l$, $s_{lm}$ — количество критических классов в подцепочке $K_l, K_{l+1}, \ldots, K_m$.



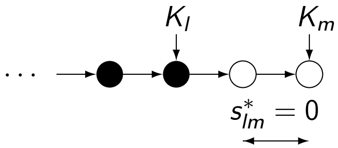
# Математическое ожидание числа применений правила

$M_{ij}$ — математическое ожидание числа применения правила $r_{ij} : A_i \xrightarrow{p_{ij}} \beta_{ij}$ в деревьях высоты $t$.

$$M_{ij}(t) = \mu_{ij} \cdot t^{1+\left(\frac{1}{2}\right)^{s^*_{lm}}} \cdot (1 + o(1))$$

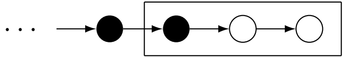



Когда $A_i$ находится в последнем критическом классе, либо дальше в цепочке, $M_{ij}(t) \sim \mu_{ij} \cdot t^2$.



# Энтропия множества деревьев высоты $t$

Наибольший вклад в $M_{ij}(t)$ вносят правила $r_{ij} : A_i \xrightarrow{p_{ij}} \beta_{ij}$, для которых $A_i$ находится в последнем критическом классе, либо классах, следующих за ним.



Вероятности таких правил определяют энтропию множества $D^t$ деревьев вывода высоты $t$.

$$H(D^t) = \sum_{\substack{A_i \in K_{h_i} \\ K_l \prec_* K_{h_i}}} d_i \sum_{j=1}^{n_i} p_{ij} \log p_{ij} \cdot t^2 \cdot (1 + o(1))$$

# Стоимость оптимального кодирования

Пусть $f$ — некоторое кодирование языка $L$.
$C(L, f)$ — стоимость кодирования $f$ языка $L$.

Тогда стоимость кодирования можно оценить

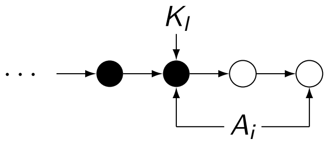
$$C(L, f) \geqslant C^*(L),$$

величиной

$$C^*(L) = \frac{\sum_{i \in I_l^+} d_i H(R_i)}{\sum_{i \in I_l^+} d_i L(R_i)},$$

где

$$I_l^+ = \{i : A_i \in K_{h_i}, K_l \preccurlyeq_* K_{h_i}\}$$



Спасибо за внимание!