Analysis of symbol statistics in bicomponent rational models

Massimiliano Goldwurm[1], Jianyi Lin[2], Marco Vignati[1]

[1]Dipartimento di Matematica, Università degli Studi di Milano, Italy

[2]Department of Mathematics, Khalifa University, Abu Dhabi, United Arab Emirates

DLT 2019, Warsaw, August 5-9, 2019

Pattern Statistics

finite alphabet A

pattern $a \in A$,
$w \in A^+$, $|w| = m$
$R \subseteq A^*$, finite, regular, ...

random text $x \in A^+$, $|x| = n$
stochastic model (Bernoullian, Markovian,.., rational)

$$O_n = \sharp\{\text{occurrences of pattern in } x\} \qquad \text{(positions)}$$

$$O_n \in \{0, 1, \ldots, n\}$$

Goals: - asymptotic properties of $\{O_n\}$
- $E(O_n)$, $var(O_n)$, ... ;
- limit distributions, $Pr\{O_n \leq z\} \sim \cdots$
- local limit laws, $Pr\{O_n = k\} \sim \cdots$

[Guibas-Odlyzko 78, 81], [Regnier-Szpankowski 96, 98],

[Nicodeme-Salvy-Flajolet 02], [Flajolet-Szpankowski-Vallée 06], ...

Rational models

[Bertoni-Choffrut-G-Lonati 03] ... [GLV18]

finite alphabet $\{a, b\}$,
pattern a,
random text fix $r : \{a, b\}^* \to \mathbb{R}_+$ rational
r defined by a $\mathbb{R}_+$-weighted non-deterministic f.s. automaton $\mathcal{A}$ over $\{a, b\}$
choose $x \in \{a, b\}^n$, with $\Pr(x) = \frac{r(x)}{\sum_{w \in \{a,b\}^n} r(w)}$

symbol statistics $(\forall n \in \mathbb{N})$
$Y_n = |x|_a$,
$\{Y_n\}$ indipendent random variables
$\Rightarrow \quad Y_n \in \{0, 1, \ldots, n\}$,
$Y_n = Y_n(r)$

Special cases: - $r = \chi_L \Rightarrow L$ is rational

$\Rightarrow x \in L \cap \{a,b\}^n$ under uniform dist.

- Markovian source (π, P) with set of states A

$$\pi' P^n e = \sum_{w \in \{a,b\}^n} r(w) = 1, \ \forall n \in \mathbb{N}$$

Regular pattern + Markovian source $\subsetneq$ Rational models

[Nicodeme-Salvy-Flajolet 02, BCGL03]

1) For every: $\begin{cases} A & \text{finite alphabet} \\ R \subseteq A^* & \text{regular language (pattern)} \\ (\pi, P) & \text{Markovian source over } A \end{cases}$

there exists $r \in \mathbb{R}_+^{\text{rat}} \langle\langle a, b \rangle\rangle$ s.t.

$O_n(\pi, P, R)$ and $Y_n(r)$ have the same distribution

i.e. $\forall k = 0, ..., n$, $\Pr(O_n = k) = \Pr(Y_n = k)$

2) The opposite is not true:

$\Longrightarrow$ Rational models are more general than Markovian models

Formal notion of rational model

Linear representation for $r \in \mathbb{R}_+^{\mathrm{rat}}\langle\langle a, b\rangle\rangle$

(ξ, A, B, η)

- $\xi, \eta \in \mathbb{R}_+^m$, weights of initial and final states ($m \in \mathbb{N}$)
- $A, B \in \mathbb{R}_+^{m\times m}$ ($\neq 0$), weights of a- and b-transitions

A, B generate a morphism $\mu : \{a, b\}^* \to \mathbb{R}_+^{m\times m}$

$$\mu(a) = A, \mu(b) = B$$

$$\mu(a_1 a_2 \cdots a_n) = \mu(a_1)\mu(a_2)\cdots\mu(a_n)$$

Rational series $r \in \mathbb{R}_+^{\mathrm{rat}}\langle\langle a, b\rangle\rangle$ defined by (ξ, A, B, η):

for every $x = a_1 a_2 \cdots a_n \in \{a, b\}^*$

$$r(x) = \xi'\mu(x)\eta = \xi'\mu(a_1)\mu(a_2)\cdots\mu(a_n)\eta$$

Probability Measure over $\{a,b\}^n$

$$\Pr(x) = \frac{\xi' \mu(w) \eta}{\xi'(A+B)^n \eta}, \quad \forall x \in \{a,b\}^n$$

Distribution of $Y_n = |x|_a$

$$\Pr(Y_n = k) = \frac{[z^k]\xi'(Az+B)^n \eta}{\xi'(A+B)^n \eta}, \quad \forall\, k = 0,1,...,n.$$

Characteristic function of Y_n

$$\Psi_n(t) = \frac{\xi'(Ae^{it}+B)^n \eta}{\xi'(A+B)^n \eta}$$

Case analysis according to $M = A + B$ [BCGL03, DGL04, BCGL06]

- M primitive $\rightarrow$ Gaussian limit dist.
- M two components $\rightarrow$ limit dist.
  - Gaussian
  - uniform
  - normal Mixture....
- M multicomponent $\rightarrow$ ....“Vandermonde” limit dist.’s

Global $\leftrightarrow$ Local limit distributions

$\{X_n\}$ r.v.'s , $X_n \in \{0, 1, \ldots, n\}$
X r.v. of distribution function $F(z) = \Pr(X \leq z)$ $(z \in \mathbb{R})$
and density function $f(z)$

1) Convergence in distribution

$$X_n \xrightarrow{d} X \text{ if}$$
$$\lim_{n \to +\infty} \Pr(X_n \leq z) = F(z) \quad \forall\, z \in \mathbb{R} \ \ (F \in C(z))$$

2) Local convergence (idea)

$\exists\, \{s_n\} \subset \mathbb{R}_+$, $z_{n,k} \subset \mathbb{R}$:

$$|s_n \Pr(X_n = k) - f(z_{n,k})| \to 0 \qquad (\text{as } n \to +\infty)$$
uniformly for $k = 0, 1, \ldots, n$

Convergence in distribution $\not\Rightarrow$ Local convergence
Additional conditions are necessary (aperiodicity)

Gaussian local limit Laws

$\{X_n\}$ r.v.'s , $X_n \in \{0, 1, \ldots, n\}$

$\{X_n\}$ satisfies a local limit law of Gaussian type

if $\exists \{a_n\}, \{s_n\}, \{\epsilon_n\} \subset \mathbb{R}, s_n > 0,$

$$a_n \sim \mathrm{E}(X_n), \quad s_n^2 \sim \mathrm{Var}(X_n), \quad \epsilon_n \to 0$$

such that

$$\left| s_n \Pr\left(X_n = k\right) - \frac{e^{-\left(\frac{k-a_n}{s_n}\right)^2/2}}{\sqrt{2\pi}} \right| \leq \epsilon_n$$

uniformly for $k \in \{0, 1, \ldots, n\}$

ϵ_n is the convergence rate

Example de Moivre - Laplace Local Limit Theorem

$\{X_{n,p}\}_n$ Binomial r.v.'s of parameter $p \in (0, 1)$ $(q = 1 - p)$

$$\left| \sqrt{2\pi npq}\, \Pr\left(X_{n,p} = k\right) - e^{-\frac{(k-np)^2}{2npq}} \right| = O(n^{-1/2})$$

Primitive rational models

[BCGL03, BCGL06]

(ξ, A, B, η) with primitive $M = A + B$, $A \neq 0 \neq B$

Main parameters: λ, β, γ

$\lambda > 0$	main eigenvalue of M
$\beta > 0$	mean constant $(0 < \beta < 1)$
$\gamma > 0$	variance constant

Results:

$$E(Y_n) = \beta n + c + o(1), \qquad c \in \mathbb{R}$$
$$var(Y_n) = \gamma n + O(1)$$

$$\frac{Y_n - \beta n}{\sqrt{\gamma n}} \longrightarrow \mathcal{N}(0,1) \qquad \text{in dist.}$$

Local limit law (in the primitive case)

Aperiodicity Condition :

G labelled graph with a-label weights A_{ij}
b-label weights B_{ij}

$$d = \mathrm{GCD}\{|C_1|_a - |C_2|_a : i \overset{C_1}{\sim\!>} i, i \overset{C_2}{\sim\!>} i, |C_1| = |C_2|\}$$

(A, B) aperiodic if $d = 1$

Theorem If M primitive, $A \neq 0 \neq B$, (A, B) aperiodic, then

$$\left| \sqrt{n} \Pr(Y_n = k) - \frac{e^{-\frac{(k-\beta n)^2}{2\gamma n}}}{\sqrt{2\pi\gamma}} \right| = O\left(n^{-1/2}\right)$$

uniformly for every $k \in \{0, 1, \ldots, n\}$.

Improves on [BCGL03]

Bicomponent models (with communication)

Model formed by two communicating irreducible components (ξ, A, B, η) of size $m_1 + m_2$ such that

$$\xi' = (\xi'_1, \xi'_2), \quad A = \left[\begin{array}{c|c} A_1 & A_0 \\ \hline 0 & A_2 \end{array}\right], \quad B = \left[\begin{array}{c|c} B_1 & B_0 \\ \hline 0 & B_2 \end{array}\right], \quad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

$(\xi_1, A_1, B_1, \eta_1)$ of size m_1, $(\xi_2, A_2, B_2, \eta_2)$ of size m_2

$$M = \left[\begin{array}{c|c} A_1 + B_1 & A_0 + B_0 \\ \hline 0 & A_2 + B_2 \end{array}\right]$$

Hypotheses:

1) $A_1 + B_1$ and $A_2 + B_2$ irreducible
with λ_1, λ_2 main eigenvalues
2) $A_0 + B_0 \neq 0$ and $\xi_1 \neq 0 \neq \eta_2$
communication cond. $1 \rightsquigarrow 2$

$\Rightarrow$ various cases depending on: $\lambda_1 \stackrel{?}{=} \lambda_2$, $\beta_1 \stackrel{?}{=} \beta_2$, $\gamma_1 \stackrel{?}{=} \gamma_2$

Dominant bicomponent models: $\lambda_1 \neq \lambda_2$

Hp: $\lambda_1 > \lambda_2$, M_1 primitive, $A_1 \neq 0 \neq B_1$ $(\Rightarrow\ 0 < \beta_1 < 1,\ 0 < \gamma_1)$

$$E(Y_n) \sim \beta_1 n, \qquad var(Y_n) \sim \gamma_1 n, \qquad \frac{Y_n - \beta_1 n}{\sqrt{\gamma_1 n}} \xrightarrow{d} \mathcal{N}(0,1)$$

[dGL04]

Theorem Under the same hps, if (A_1, B_1) aperiodic then

$$\left| \sqrt{n}\Pr(Y_n = k) - \frac{e^{-\frac{(k-\beta_1 n)^2}{2\gamma_1 n}}}{\sqrt{2\pi\gamma_1}} \right| = O\left(n^{-1/2}\right)$$

uniformly for every $k \in \{0, 1, \ldots, n\}$.

Equipotent bicomponent models with different β's

Hp: $\lambda_1 = \lambda_2$, $\beta_1 \neq \beta_2$,
M_1, M_2 primitive, $A_j \neq 0 \neq B_j$ $(j = 1, 2)$

$$E(Y_n) = \frac{\beta_1 + \beta_2}{2} n + O(1),$$

$$var(Y_n) = \frac{(\beta_1 - \beta_2)^2}{12} n^2 + O(n)$$

$$\frac{Y_n}{n} \xrightarrow{d} U \quad \text{(convergence in dist. to a uniform r.v.)}$$

$$\text{where } f_U(x) = \begin{cases} 0 & \text{if } x \leq b_1 \\ \frac{1}{b_2 - b_1} & \text{if } b_1 < x \leq b_2 \\ 0 & \text{if } b_2 < x \end{cases}$$

$$\text{with } b_1 = \min\{\beta_1, \beta_2\}, b_2 = \max\{\beta_1, \beta_2\}$$

Local limit law of uniform type

$$f_U(x) = \begin{cases} 0 & \text{if } x \le b_1 \\ \frac{1}{b_2 - b_1} & \text{if } b_1 < x \le b_2 \\ 0 & \text{if } b_2 < x \end{cases}, \quad \text{where } \begin{cases} b_1 = \min\{\beta_1, \beta_2\}, \\ b_2 = \max\{\beta_1, \beta_2\} \end{cases}$$

Theorem

Under the same hps $\begin{cases} \lambda_1 = \lambda_2, \beta_1 \neq \beta_2 \\ M_1, M_2 \text{ primitive}, A_j \neq 0 \neq B_j,\ j = 1, 2 \end{cases}$

if $(A_1, B_1), (A_2, B_2)$ aperiodic then

$$|n \Pr(Y_n = k) - f_U(x)| = O\left(\frac{(\log n)^{3/2}\, \tau_n}{\sqrt{n}}\right)$$

for all $k = k(n) \in [n]$ s.t. $k/n \to x$, $x \in \mathbb{R}$, $\beta_1 \neq x \neq \beta_2$;
where $\{\tau_n\} \subset \mathbb{R} : \tau_n \to +\infty$, $\tau_n = o(\log \log n)$ (arbitrarily slow)

Example

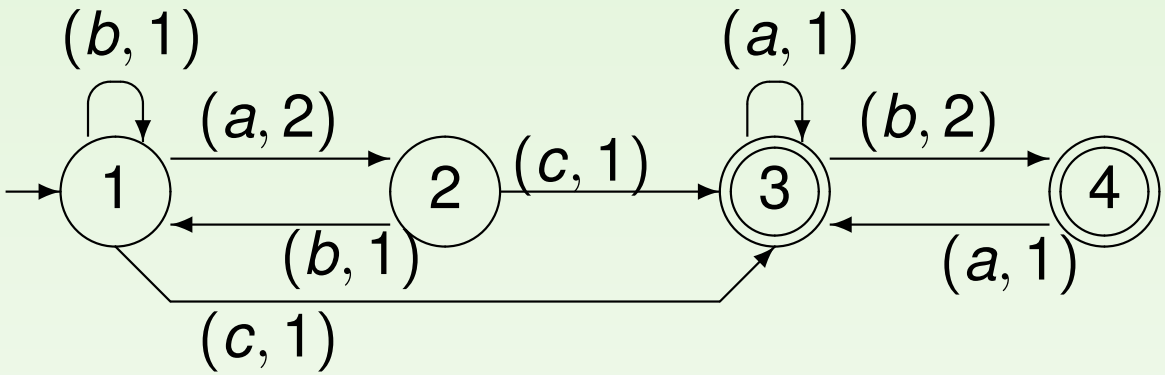

Figure: $\lambda_1 = \lambda_2 = 2$, $1/3 = \beta_1 \neq \beta_2 = 2/3$ (equipotent)

$L = \{x \in \{a, b\}^* \mid aa \notin x\}\ c\ \{y \in \{a, b\}^* \mid bb \notin y\}$
$M_1 = M_2$, $A_1 \neq A_2$, $(A_1, B_1), (A_2, B_2)$ aperiodic

$\Rightarrow$ Y_n/n has local limit of uniform type

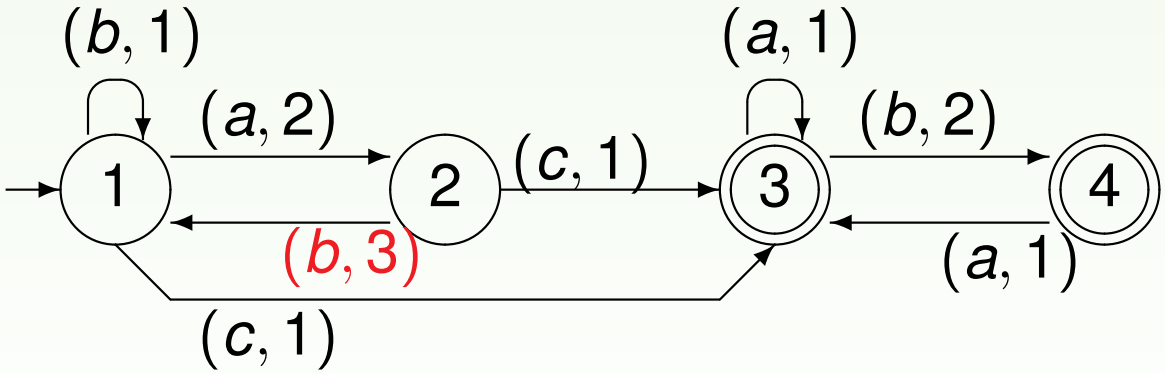

Figure: $\lambda_1 = 3, \lambda_2 = 2$ $\Rightarrow$ Y_n/n has local limit of Gaussian type

Equipotent bicomponent models with equal β's and different γ's

Hp: $\lambda_1 = \lambda_2$, $\beta_1 = \beta_2$, $\gamma_1 \neq \gamma_2$,
M_1, M_2 primitive, $A_j \neq 0 \neq B_j$ $(j = 1, 2)$

Set: $\beta = \beta_1 = \beta_2$, $\gamma = \frac{\gamma_1 + \gamma_2}{2}$
r.v. T mixture of $\mathcal{N}(0, s)$ with
variance s uniformly distributed between $\frac{\gamma_1}{\gamma}$ and $\frac{\gamma_2}{\gamma}$.

$$f_T(x) = \frac{\gamma}{\gamma_2 - \gamma_1} \int_{\frac{\gamma_1}{\gamma}}^{\frac{\gamma_2}{\gamma}} \frac{e^{-\frac{x^2}{2s}}}{\sqrt{2\pi s}}\, ds \qquad \forall\, x \in \mathbb{R}$$

$f_T(x)$ $\rightsquigarrow$ heat equation in dimension 1

It is known that

$$\frac{Y_n - \beta n}{\sqrt{\gamma n}} \xrightarrow{d} T$$

Local limit law of T type

Theorem

Under the same hps $\begin{cases} \lambda_1 = \lambda_2, \beta_1 = \beta_2, \gamma_1 \neq \gamma_2 \\ M_1, M_2 \text{ primitive}, A_j \neq 0 \neq B_j,\ j = 1, 2 \end{cases}$

if $(A_1, B_1), (A_2, B_2)$ aperiodic and $\gamma = \frac{\gamma_1 + \gamma_2}{2}$ then

$$\left| \sqrt{\gamma n} \Pr(Y_n = k) \ - \ f_T \left(\frac{k - \beta n}{\sqrt{\gamma n}} \right) \right| \ = \ O\left(\frac{(\log n)^2 \ \tau_n}{\sqrt{n}} \right)$$

uniformly for $k \in \{0, 1, \ldots, n\}$

where $\{\tau_n\} \subset \mathbb{R} : \tau_n \to +\infty, \tau_n = o(\log \log n)$ (arbitrarily slow)

Equipotent bicomponent models with equal β's and γ's

Theorem

Assume the bicomponent model,
let $\lambda_1 = \lambda_2$, $\beta_1 = \beta_2$, $\gamma_1 = \gamma_2$
M_1, M_2 primitive, $A_j \neq 0 \neq B_j$ for $j = 1, 2$.
If (A_1, B_1), (A_2, B_2) are aperiodic then

$$\left| \sqrt{n} \Pr(Y_n = k) - \frac{e^{-\frac{(k-\beta n)^2}{2\gamma n}}}{\sqrt{2\pi\gamma}} \right| = O\left(n^{-1/2}\right)$$

uniformly for every $k \in \{0, 1, \ldots, n\}$.

Summary of results

	Primitive Models	Bicomponent Models			
			$\lambda_1 = \lambda_2$		
		$\lambda_1 \neq \lambda_2$	$\beta_1 \neq \beta_2$	$\beta_1 = \beta_2$ $\gamma_1 \neq \gamma_2$	$\beta_1 = \beta_2$ $\gamma_1 = \gamma_2$
Local limit distribution	$N_{0,1}$ [BCGL03]	$N_{0,1}$	U_{β_1,β_2} [GLV18]	T	$N_{0,1}$
Convergence rate	$O(n^{-1/2})$	$O(n^{-1/2})$	$O\left(\frac{\tau_n \log^{3/2} n}{\sqrt{n}}\right)$	$O\left(\frac{\tau_n \log^2 n}{\sqrt{n}}\right)$	$O(n^{-1/2})$

Conclusions

- The results strengthen previous convergence in distribution [BCGL03],[dGL04]
- Convergence rate $O(n^{-1/2})$ for all Gaussian limits;
- Convergence rate a bit "slower" than $O(n^{-1/2})$ otherwise;
- Examples of non-Gaussian local limit laws;
- Proofs $\longrightarrow$ Saddle Point method;
- Analysis of bicomponent models without communication [GLV19b]

  $$A_0 + B_0 = 0, \quad (r = s + t \text{ for } s, t \in \mathbb{R}_+^{\text{rat}}\langle\langle a, b\rangle\rangle)$$

  $\Rightarrow$ different limit distributions with $O(n^{-1/2})$ convergence rate
- Open problems:
  - why different convergence rates?
  - analysis of multicomponent cases
  - more general probabilistic models
  - different notion of pattern occurrences

Thank you !