S©MR

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ
Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru
ISSN 1813-3304

Vol. 20, No. 2, pp. 1–11 (2025) УДК 519.21
https://doi.org/10.33048/semi.2025.20.??? MSC 60J10

ASYMPTOTICS OF THE NUMBER OF DIFFERENT WORDS IN A MARKOV CHAIN DRIVEN MODEL

SH.SH.FAYZULLAEV AND **A.P. KOVALEVSKII**

Abstract: This paper investigates the asymptotic of the number of distinct words in a finite Markov chain driven model. We analyse the normalized and centered processes associated with the occurrence of distinct words in the model. Each state of the Markov chain is associated with its own unique infinite dictionary. At each state of the Markov chain, words are selected from the dictionary according to an infinite urn scheme. The probabilities in each infinite urn scheme satisfy the condition of regular variation. We use a combination of asymptotic techniques and results for Gaussian processes and derive the covariance structure of the limiting processes. The influence of stationary probabilities of the Markov chain on the normalization and scaling of these processes is explored in detail. Our findings provide new insights into the interaction between word frequencies and the stationary distribution in systems with overlapping or non-overlapping dictionaries. These results are applicable

Fayzullaev Sh.Sh., Kovalevskii A.P., Asymptotics of the number of different words in a Markov chain driven model.

The research of Shahzod Fayzullaev is supported by the "El-Yurt Umidi" Foundation for Training of Promising Personnel under the President of the Republic of Uzbekistan.

The work is supported partially by the Fundamental scientific research of the SB RAS, project FWNF-2022-0010.

Received January, 1, 2025, Published December, 31, 2025.

to a wide range of stochastic systems, offering a deeper understanding of their limiting behaviour.

Keywords: Stationary distribution, Markov processes, Different words, Infinite urn scheme.

1 Introduction

Karlin [1] investigated an infinite-box scheme in which each ball is independently placed in the ith box with probability p_i. In particular, Karlin studied the asymptotics of the number of non-empty boxes after n throws and proved the Central Limit Theorem for this statistic. This infinite box scheme can be interpreted as a fundamental elementary probabilistic model for text generation: the boxes correspond to the words of an infinite dictionary, while the balls represent the sequential words in a text. Within this framework, words are selected randomly and independently. This model has been studied in many works over the years. The first explanation of the Law of Large Numbers for distinct values in the elementary probabilistic model of the infinite urn scheme is found in [10]. Before addressing a few of them, let's discuss [2], which forms the basis of our main work. In this research, a new "regime-switching model"was proposed. The essence of the model is as follows: let A be a finite set, and $X = (X_n)_{n\geq 1}$ be a discrete Markov chain taking values in the set A. In each state $a \in A$, items are taken independently and identically distributed (iid) with distribution p_a from an infinite set (vocabulary) V_a. The vocabularies do not overlap, $V_a \cap V_b = \emptyset$ for $a \neq b$. We assume that the Markov chain $\{X_i\}$ is irreducible and starts from its stationary distribution $\{\pi_a, a \in A\}$. Pananjady et al. [8] proposed a new method for estimating the missing mass in Markov chains, called Windowed Good-Turing (WingIt). The method is minimax-optimal in terms of mean squared error and has linear runtime complexity, making it efficient for large-scale data. They also conducted experiments demonstrating the effectiveness of their approach on synthetic data and natural language texts.

If $A = \{a\}$ has only one element, then we have the classical results of Karlin. Random variables $Y_1^a, Y_2^a, \ldots$ are iid and take values in a countable vocabulary $V_a = \{1_a, 2_a, \ldots\}$. Let

$$p_k^a := \mathbb{P}(Y_1^a = k) = l(a,k)k^{-\alpha_a}, \quad k \in V_a, \tag{1}$$

where $\alpha_a > 1$ is an unknown parameter, and $l(a,k) > 0$ is a function satisfying the normalization condition:

$$\sum_{k\in V_a} l(a,k)k^{-\alpha_a} = 1.$$

The function $l(a,x)$ is slowly varying as $x \to \infty$ for fixed a, meaning that:

$$\frac{l(a,cx)}{l(a,x)} \to 1 \quad \text{as } x \to \infty \text{ for any } c > 0.$$

Let $J_i^a(n)$ denote the number of balls in urn i after drowing n balls in state a,

$$J_i^a(n) = \sum_{j=1}^{n} \mathbf{1}(Y_j^a = i).$$

Let R_n^a be the number of nonempty urns and $R_{n,k}^{*,a}$ be the number of urns containing at least $k \geq 1$ balls. These quantities can be expressed as:

$$R_n^a = \sum_{i=1}^{\infty} \mathbf{1}(J_i^a(n) > 0), \quad R_{n,k}^{*,a} = \sum_{i=1}^{\infty} \mathbf{1}(J_i^a(n) \geq k),$$

where $\mathbf{1}(\cdot)$ be the indicator function.

Notably, $R_{n,1}^{*,a} = R_n^a$. Furthermore, the number of urns containing exactly k balls is given by:

$$R_{n,k}^a = R_{n,k}^{*,a} - R_{n,k+1}^{*,a}.$$

Khmaladze [9] investigates asymptotic properties of occupancy statistics, including the Karlin–Rowault law, which describes the distribution of the number of bins with a given occupancy rate. In particular, the author showed that for a system with n objects and N bins, the ratio $\mathbb{E}R_{n,k}^a / \mathbb{E}R_n^a$ converges to a function $\rho_{\theta_a}(k)$, defined as

$$\rho_{\theta_a}(k) = \frac{\theta_a \Gamma(k - \theta_a)}{\Gamma(k+1)\Gamma(1-\theta_a)},$$

where $\theta_a = 1/\alpha_a$.

Karlin [1] proposed studying a random sample where the number of experiments is itself random, denoted by $\Pi_a(n)$. Here, $\{\Pi_a(t), t \geq 0\}$ represents a Poisson process with parameter 1. The random choice of an urn and the Poisson process are assumed to be independent. From the splitting property of the Poisson process, the processes $\{J_i^a(\Pi_a(t)) = \Pi_i^a(t), t \geq 0\}$ are independent Poisson processes with respective parameters p_i^a.

By definition, the quantity $R_{\Pi(t),k}^{*,a}$ is expressed as:

$$R_{\Pi_a(t),k}^{*,a} = \sum_{i=1}^{\infty} \mathbf{1}(\Pi_i^a(t) \geq k).$$

In [16] the asymptotics of the number of unique words was studied. In particular, it was shown that if $\lim_{i\to\infty} p_{i+1}^a / p_i^a = 1$, then $R_{n,1}^a \to_p \infty$ for $n \to \infty$. If $\limsup_{i\to\infty} p_{i+1}^a / p_i^a < 1$, then $\mathbf{E}R_{n,1}^a$ remains uniformly bounded. The statistics R_n^a and $R_{n,1}^a$ in these models are also studied in the works of [3], [4], [5], [6], [7], [12], [13], [14], [15], [17]. We present to your attention [20] about statistical tests conducted using these statistics.

In [18], a randomized Karlin scheme with parameter $\beta \in (0,1)$ is studied. It is shown that under restricted randomization the odd-occupancy process after normalization scales to fractional Brownian motion with Hurst index $\beta/2 \in (0, \frac{1}{2})$, while under heavy tails of randomization distribution with index $\alpha \in (0,2)$ the same process converges to (β/α)-self-similar symmetric α-stable process with stationary increments.

Another new work (see [19]) introduces weighted processes $\sum_j w_j R^a_{n,j}$ with flexible weights w_j. Under the condition $|w_{i+j} - w_i| \leq Cj^\beta$, a Functional CLT is proved.

We refer to [22] for modifications of the Karlin (also Simon, see [21]) text model.

1.1. Related work.

We define $\theta_a = 1/\alpha_a$, where θ_a represents the inverse Zipf exponent, satisfying $0 < \theta_a < 1$.

A general formulation of (1), as presented in [1], is given by:

$$\kappa_a(x) := \max\{k > 0 : \ p^a_k \geq 1/x\} = L(a,x)x^{\theta_a}, \tag{2}$$

where $L(a,x)$ is a slowly varying function. Specifically, for any $c > 0$, the property

$$\frac{L(a,cx)}{L(a,x)} \to 1 \quad \text{as } x \to \infty$$

holds.

Karlin established the almost sure (a.s.) convergences:

$$\frac{R^a_n}{\mathbb{E}R^a_n} \to 1 \quad \text{and} \quad \frac{R^a_{n,1}}{\mathbb{E}R^a_{n,1}} \to 1.$$

In [11] was proved that there is convergence of the centered and normalized process of numbers of different words,

$$Z^a_n := \{Z^a_n(t),\ 0 \leq t \leq 1\ \} = \left\{ \frac{R^a_{[nt]} - \mathbb{E}R^a_{[nt]}}{\sqrt{\mathbb{E}R^a_n}},\ 0 \leq t \leq 1 \right\} \tag{3}$$

converges weakly in $D(0,1)$ with uniform metrics to a centered Gaussian process Z with continuous a.s. sample paths and covariance function

$$K_a(t,s) = (s+t)^{\theta_a} - \max(s^{\theta_a}, t^{\theta_a}). \tag{4}$$

Karlin [1] proved that

$$\mathbb{E}R^a_n \sim \Gamma(1-\theta_a)L(a,x)n^{\theta_a}, \quad \mathbb{E}R^a_{n,k} \sim \theta_a \frac{\Gamma(k-\theta_a)}{k!} L(a,x)n^{\theta_a}. \tag{5}$$

2 Main results

Let $A = \{a_1, a_2, \ldots, a_{|A|}\}$ be a finite set of states of an irreducible Markov chain $\{X_i\}_{i\geq 0}$. Assume the Markov chain starts in the stationary regime:

$$\mathbf{P}(X_i = a_j) = \pi_j, \quad i \geq 0,\ j \leq |A|.$$

We have the CLT for the Markov chain. Denote the sum $S_n = X_1 + \cdots + X_n$. Consequently, at this time the weak invariance principle holds; in other words, putting

$$\widetilde{S}_n(t) = \frac{S_{[nt]} - \mathbb{E}S_{[nt]}}{\sqrt{\mathrm{Var}(S_n)}}, \quad 0 \le t \le 1, \tag{6}$$

we have

$$\widetilde{S}_n \Rightarrow W \quad \text{as } n \to \infty, \tag{7}$$

where $\{W(t),\, 0 \le t \le 1\}$ is a standard Wiener process on $D([0,1])$ with Skorokhod topology.

Denote $N_{a,n} = \sum_{i=1}^{n} \mathbf{1}(X_i = a)$, $a \in A$.

It is obvious that $\mathbb{E}N_{a,n} = n\pi_a$.

Let

$$\widetilde{N}_{a,n}(t) = \frac{N_{a,[nt]} - \mathbb{E}N_{a,[nt]}}{\sqrt{\mathrm{Var}(N_{a,n})}}, \quad 0 \le t \le 1, \tag{8}$$

From the Functional CLT for $(N_{a,n},\ a \in A)$ we have

$$(\widetilde{N}_{a,n},\ a \in A) \Rightarrow (W_a,\ a \in A) \quad \text{as } n \to \infty$$

on $D^{|A|}([0,1])$ with Skorokhod topology, components W_a are standard Wiener processes,

$$\mathbf{Cov}(W_a(s), W_b(t)) = \lim_{n\to\infty} \mathbf{Cov}\left(\frac{N_{a,[ns]} - \mathbb{E}N_{a,[ns]}}{\sqrt{\mathrm{Var}(N_{a,n})}}, \frac{N_{b,[nt]} - \mathbb{E}N_{b,[nt]}}{\sqrt{\mathrm{Var}(N_{b,n})}} \right)$$

$$= \lim_{n\to\infty} \frac{\mathbf{Cov}(N_{a,[ns]}, N_{b,[nt]})}{\sqrt{\mathrm{Var}(N_{a,n})}\sqrt{\mathrm{Var}(N_{b,n})}}$$

This limit can be calculated as follows: let's start with

$$\mathrm{Cov}(N_{a,n}, N_{b,n}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{Cov}(\mathbf{1}(X_i = a), \mathbf{1}(X_j = b))$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{a,b}(j-i) = |j - i = k| = \sum_{k=-(n-1)}^{n-1} (n - |k|)\gamma_{a,b}(k),$$

where

$$\gamma_{a,b}(k) := \mathbf{Cov}(\mathbf{1}(X_0 = a), \mathbf{1}(X_k = b)), \quad k \in \mathbb{Z}.$$

For $k = 0$:

$$\gamma_{a,b}(0) = \begin{cases} \pi_a(1-\pi_a), & \text{if } a = b, \\ -\pi_a\pi_b, & \text{if } a \ne b. \end{cases}$$

For $k > 0$:

$$\gamma_{a,b}(k) = \mathbb{P}(X_0 = a, X_k = b) - \pi_a\pi_b = \pi_a P^k(a,b) - \pi_a\pi_b.$$

For $k < 0$:

$$\gamma_{a,b}(k) = \gamma_{b,a}(-k) = \pi_b P^{-k}(b,a) - \pi_a\pi_b.$$

$$\frac{\mathbf{Cov}(N_{a,n}, N_{b,n})}{n} = \sum_{k=-(n-1)}^{n-1} \left(1 - \frac{|k|}{n}\right)\gamma_{a,b}(k) \to \sum_{k=-\infty}^{\infty} \gamma_{a,b}(k) =: v_{a,b}.$$

Similarly,

$$\mathrm{Var}(N_{a,n}) = \sum_{i=1}^{n} \mathrm{Var}(1(\mathrm{X_i = a})) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{Cov}(1(\mathrm{X_i = a}), 1(\mathrm{X_j = a})) =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \gamma_{a,a}(j-i) = \sum_{k=-(n-1)}^{n-1} (n-|k|)\gamma_{a,a}(k).$$

From here

$$\frac{\mathrm{Var}(N_{a,n})}{n} \to v_{a,a}, \quad a \in A.$$

Hence,

$$\lim_{n\to\infty} \frac{\mathbf{Cov}(N_{a,n}, N_{b,n})}{\sqrt{\mathrm{Var}(N_{a,n})}\sqrt{\mathrm{Var}(N_{b,n})}} = \frac{v_{a,b}}{\sqrt{v_{a,a}}\sqrt{v_{b,b}}} =: \rho_{a,b}.$$

$$\mathbf{cov}(W_a(s), W_b(t)) = \lim_{n\to\infty} \frac{\mathbf{Cov}(N_{a,[ns]}, N_{b,[nt]})}{\sqrt{\mathrm{Var}(N_{a,n})}\sqrt{\mathrm{Var}(N_{b,n})}} = \min(s,t)\rho_{a,b}$$

Let each of $a \in A$ corresponds to its own dictionary.

Let (2) be satisfied for any state $a \in A$. Without loss of generality, we rearrange the set of states A so that their parameters are arranged in descending order:

$$\theta_{a_1} \geq \theta_{a_2} \geq \cdots \geq \theta_{a_{|A|}}. \tag{9}$$

Let k_0 be such that

$$\theta_{a_1} = \theta_{a_2} = \ldots = \theta_{a_{k_0}} > \theta_{a_{k_0+1}} \geq \cdots \geq \theta_{a_{|A|}}, \tag{10}$$

$1 \leq k_0 \leq |A|$.

Let there exist

$$c_j := \lim_{x\to\infty} L(a_j, x)/L(a_1, x) \leq 1, \quad 1 \leq j \leq k_0. \tag{11}$$

Note that $c_1 = 1$.

Let $R^a_{N_{a,n}}$ denote the number of distinct words observed in state a, and $R^a_{N_{a,n},1}$ denotes the number of words that occur only once in state a. While the transitions between states follow a Markov chain with stationary distribution $\pi = (\pi_a)_{a\in A}$. Since the dictionaries corresponding to different states $a \in A$ do not intersect, the number of distinct words $\mathcal{R}_n$ and the number of words that occur only once $\mathcal{R}_{n,1}$ are determined by the sums $\mathcal{R}_n = \sum_{a\in A} R^a_{N_{a,n}}$ and $\mathcal{R}_{n,1} = \sum_{a\in A} R^a_{N_{a,n},1}$, respectively. From $N_{a,n} = \mathbb{E}N_{a,n} + O_P(\sqrt{\mathrm{Var}N_{a,n}})$ we now $\mathbb{E}R^a_{N_{a,n}} \sim \mathbb{E}R^a_{[\mathbb{E}N_{a,n}]} = \mathbb{E}R^a_{[n\pi_a]}$. Let $R^a_{\Pi_a(N_{a,n}),1}$ and $R^{a,*}_{\Pi_a(N_{a,n}),1}$ denote the number of words in state a that occur only once and the number of distinct words, respectively, after n steps of the Markov chain under Poissonization. Similarly, for a Poisson setting, the total number of words that occur only once and the total number of distinct words at time n can be expressed as $\mathcal{R}^{\Pi}_{n,1} = \sum_{a\in A} R^a_{\Pi_a(N_{a,n}),1}$ and $\mathcal{R}^{\Pi}_n = \sum_{a\in A} R^{a,*}_{\Pi_a(N_{a,n}),1}$, respectively. In this model, Poissonization controls the random occurrence of words in each state $a \in A$, where transitions between states follow a Markov chain with stationary distribution $\pi = (\pi_a)_{a\in A}$. Observations in each state are independent and identically distributed according to a regime-specific distribution $p_{a,m}$. Poisson processes $\{\Pi_a(\cdot), a \in A\}$ are mutually independent and does not depend on the processes of observations in each state and on the Markov chain.

We can directly derive the following asymptotics using the asymptotic formulas for the case of $|A| = 1$ from [1]:

Lemma 1. *For any $a \in A$*

$$\mathbb{E}[R^a_{\Pi_a(N_{a,n}),1}] \sim \theta_a \,\Gamma(1-\theta_a)\,(n\pi_a)^{\theta_a} L(a, n\pi_a), \tag{12}$$

$$\mathbb{E}[R^{a,*}_{\Pi_a(N_{a,n}),1}] \sim \Gamma(1-\theta_a)\,(n\pi_a)^{\theta_a} L(a, t\pi_a). \tag{13}$$

Proof. If at each step the state $a \in A$ is selected with probability π_a, then the *average* number of terms falling exactly on the label a is equivalent to $\pi_a n$. Consequently, for the number of occurrences of exactly one event at the label a by the moment n we have

$$\mathbb{E}\Big[R^a_{N_{a,n},1}\Big] \sim \theta_a\Gamma(1-\theta_a)\big(\pi_a n\big)^{\theta_a} L(a, n\pi_a),$$

When passing to Poissonization, we actually substitute $\Pi(t) \sim_{a.s.} t$ as $t \to \infty$. Then we get

$$\mathbb{E}\Big[R^a_{\Pi(N_{a,n}),1}\Big] \sim \theta_a\Gamma(1-\theta_a)\big(n\,\pi_a\big)^{\theta_a} L(a, n\pi_a).$$

Similarly, we can also obtain the (13) asymptotics.

□

Then for the entire system, the following statements are true:

Theorem 1. *Under assumptions (12) and (13),*

$$\frac{\mathbb{E}[\mathcal{R}^{\Pi}_{n,1}]}{\mathbb{E}[\mathcal{R}^{\Pi}_{n}]} \to \theta_{a_1},$$

and for any $0 < x < 1$,

$$\frac{\mathbb{E}[\mathcal{R}^{\Pi}_{[xn]}]}{\mathbb{E}[\mathcal{R}^{\Pi}_{n}]} \to x^{\theta_{a_1}}, \qquad \frac{\mathbb{E}[\mathcal{R}^{\Pi}_{[xn],1}]}{\mathbb{E}[\mathcal{R}^{\Pi}_{n,1}]} \to x^{\theta_{a_1}}.$$

Proof. 1. From Lemma 1 and definitions of $\mathcal{R}^{\Pi}_{n,1}$ and $\mathcal{R}^{\Pi}_{n}$,

$$\mathbb{E}[\mathcal{R}^{\Pi}_{n,1}] = \sum_{a\in A} \mathbb{E}[R^a_{\Pi(n),1}] \sim \sum_{a\in A} \theta_a\,\Gamma(1-\theta_a)\,(n\pi_a)^{\theta_a} L(a, n\pi_a). \tag{14}$$

$$\mathbb{E}[\mathcal{R}^{\Pi}_{n}] \sim \sum_{a\in A} \Gamma(1-\theta_a)\,(n\pi_a)^{\theta_a} L(a, n\pi_a). \tag{15}$$

From (14), (15), (10) and the definition of c_j,

$$\mathbb{E}[\mathcal{R}^{\Pi}_{n,1}] \sim \theta_{a_1}\,\Gamma(1-\theta_{a_1})\,n^{\theta_{a_1}} L(a_1, n) \sum_{j\le k_0} c_j \pi_{a_j}^{\theta_{a_1}},$$

$$\mathbb{E}[\mathcal{R}_n^{\Pi}] \sim \Gamma(1-\theta_{a_1})\, n^{\theta_{a_1}} L(a_1, n) \sum_{j\le k_0} c_j \pi_{a_j}^{\theta_{a_1}}, \tag{16}$$

so

$$\frac{\mathbb{E}[\mathcal{R}_{n,1}^{\Pi}]}{\mathbb{E}[\mathcal{R}_n^{\Pi}]} \to \theta_{a_1},$$

Suppose that, according to (9), several states share a maximal parameter, i.e.

$$\theta_{a_1} = \theta_{a_2} = \cdots = \theta_{a_k} \; > \; \theta_{a_{k+1}} \ge \cdots \ge \theta_{a_{|A|}}.$$

Then the asymptotic expansions are dominated only by the terms corresponding to these maximal θ_{a_i}. Although each such term contributes to a constant, these constants appear in both the numerator and the denominator, causing them to cancel. Hence, the limit remains determined by the common maximal parameter θ_{a_1}.

Moreover, if all θ_{a_i} coincide, then the same reasoning applies: each term has the same exponent θ_{a_1}, and the overall limit again converges to θ_{a_1}.

2. Let replace n with $[xn]$ in the asymptotic expression for $\mathbb{E}[\mathcal{R}_n^{\Pi}]$:

$$\mathbb{E}[\mathcal{R}_{[xn]}^{\Pi}] = \sum_{a\in A} \Gamma(1-\theta_a)\,(xn\pi_a)^{\theta_a} L(a, xn\pi_a). \tag{17}$$

From (15) and (17) and taking into account the above reasoning, we have:

$$\frac{\mathbb{E}[\mathcal{R}_{[xn]}^{\Pi}]}{\mathbb{E}[\mathcal{R}_n^{\Pi}]} \to x^{\theta_{a_1}}.$$

Similarly for the number of words occurring only once. □

The next result is directly related to how the (3) process behaves.

Theorem 2. *Let*

$$\mathcal{Z}_n := \{\mathcal{Z}_n(t),\ 0\le t\le 1\ \} = \left\{ \frac{\mathcal{R}_{[nt]} - \mathbb{E}\mathcal{R}_{[nt]}}{\sqrt{\mathbb{E}\mathcal{R}_n}},\ 0\le t\le 1\ \right\}$$

be the normalized and centered process in the model governed by the Markov chain with a stationary distribution, and assumptions (2), (10), (11) to be hold.

Then $\mathcal{Z}_n$ converges weakly in the space $D([0,1])$ to a centered Gaussian process $Z^{a_1} = \{Z^{a_1}(t), t\in[0,1]\}$ with the covariance function given by (4) with $a = a_1$.

Proof. Consider the Poissonized version of the process,

$$\mathcal{Z}_n^{\Pi} := \{\mathcal{Z}_n^{\Pi}(t),\ 0\le t\le 1\ \} = \left\{ \frac{\mathcal{R}_{[nt]}^{\Pi} - \mathbb{E}\mathcal{R}_{[nt]}^{\Pi}}{\sqrt{\mathbb{E}\mathcal{R}_n^{\Pi}}},\ 0\le t\le 1\ \right\}$$

Its covariance function

$$\mathbf{Cov}(\mathcal{Z}_n^{\Pi}(s), \mathcal{Z}_n^{\Pi}(t)) = \frac{\mathbf{Cov}(\mathcal{R}_{[ns]}^{\Pi}, \mathcal{R}_{[nt]}^{\Pi})}{\mathbb{E}\mathcal{R}_n^{\Pi}}. \tag{18}$$

From the definition of $\mathcal{R}_n^\Pi$,

$$\mathbf{Cov}(\mathcal{R}_{[ns]}^\Pi, \mathcal{R}_{[nt]}^\Pi) = \sum_{a,b\in A} \mathbf{Cov}(R^a_{\Pi_a(N_{a,[ns]})}, R^b_{\Pi_b(N_{b,[nt]})}).$$

Poisson processes Π_a and Π_b are independent for $a \neq b$, so the the corresponding covariances are zero,

$$\mathbf{Cov}(\mathcal{R}_{[ns]}^\Pi, \mathcal{R}_{[nt]}^\Pi) = \sum_{a\in A} \mathbf{Cov}(R^a_{\Pi_a(N_{a,[ns]})}, R^a_{\Pi_a(N_{a,[nt]})}). \tag{19}$$

We know that

$$\frac{\mathbf{Cov}(R^a_{\Pi_a(ns)}, R^a_{\Pi_a(nt)})}{\mathbb{E}R^a_{\Pi_a(n)}} \to K_a(s,t) \tag{20}$$

as $n \to \infty$ for any $a \in A$.

Remember that $N_{a,n} \to \infty$ a.s. as $n \to \infty$, and $N_{a,[nt]}/N_{a,n} \to t$ a.s. by the SLLN for Markov chains. From (20),

$$\frac{\mathbf{Cov}(R^a_{\Pi_a(N_{a,[ns]})}, R^a_{\Pi_a(N_{a,[nt]})})}{\mathbb{E}R^a_{\Pi_a(N_{a,n})}} \to K_a(s,t). \tag{21}$$

Substituting (16), (19) and (20) to (18), we have

$$\mathbf{Cov}(\mathcal{Z}_n^\Pi(s), \mathcal{Z}_n^\Pi(t)) = \sum_{a\in A} \frac{\mathbf{Cov}(R^a_{\Pi_a(N_{a,[ns]})}, R^a_{\Pi_a(N_{a,[nt]})})}{\mathbb{E}R^a_{\Pi_a(N_{a,n})}} \frac{\mathbb{E}R^a_{\Pi_a(N_{a,n})}}{\mathbb{E}\mathcal{R}_n^\Pi}$$

$$\to \sum_{j\le k_0} K_{a_1}(s,t) \frac{c_j \pi_{a_j}^{\theta_1}}{\sum_{i\le k_0} c_i \pi_{a_i}^{\theta_1}} = K_{a_1}(s,t).$$

Thus, the covariances of the process converge to the covariances of the limiting Gaussian process. The convergence of multivariate distributions is ensured by the central limit theorem: since the process values are sums of independent indicators, the Lindeberg condition is satisfied for them and w weak convergence to the multivariate normal distribution is guaranteed.

The relative compactness of the distributions of processes $\{\mathcal{Z}_n^\Pi, n \ge 1\}$ follows from the fact that they are weighted sums of processes $Z^a_{\Pi_a(n)}$ on random intervals, and these random intervals converge to deterministic ones after scaling, and the relative compactness for processes on deterministic intervals is proved in Step 3 of the proof of Theorem 1 in [11]. Therefore, the Poissonized process weakly converges in the uniform metric to the limit Gaussian process.

The final step of the proof is the transition from the Poissonized version $\mathcal{Z}_n^\Pi$ of the process to the original version $\mathcal{Z}_n$. The convergence of the maximal absolute value of the difference of the processes in probability to zero is also based on the representation as a weighted sum of differences $Z^a_{\Pi_a(n)}(\cdot) - Z^a_n(\cdot)$. The convergence of each maximal absolute value of the difference in probability to zero is proved in Step 4 of the proof of Theorem 1 in [11].

Thus, process $\mathcal{Z}_n$ converges in distribution in $D([0,1])$ to a centered Gaussian process with covariance $K_{a_1}(t,s)$.

□

3 Conclusion

The process of numbers of different words, governed by a stationary indecomposable finite Markov chain, has the same limiting distribution as the one-state process. Therefore, it is not possible to obtain a model in this way that better explains the observed data than Karlin's original scheme.

Acknowledgement
The research of Shahzod Fayzullaev is supported by the "El-Yurt Umidi" Foundation for Training of Promising Personnel under the President of the Republic of Uzbekistan. The work of Artyom Kovalevskii is supported partially by the Fundamental scientific research of the SB RAS, project FWNF-2022-0010.

References

[1] S. Karlin, *Central Limit Theorems for Certain Infinite Urn Schemes.* Jounal of Mathematics and Mechanics, **17**, No. 4 ,(1967), 373–401.

[2] M. Grabchak, M. Kelbert, Q. Paris, *On the occupancy problem for a regime switching model.* *Journal of Applied Probability* **57**, No. 1,(2020), 53–77 999999

[3] A.D. Barbour, *Univariate approximations in the infinite occupancy scheme*, Alea **6** (2009), 415–433.

[4] A.D. Barbour, A.V. Gnedin, *Small counts in the infinite occupancy scheme*, Electronic Journal of Probability, Vol. 14, Paper no. 13 (2009), 365–384.

[5] A. Ben-Hamou, S. Boucheron, M. I. Ohannessian, *Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications*, Bernoulli **23**, Number 1 (2017), 249–287.

[6] M.G. Chebunin, *Estimation of parameters of probabilistic models which is based on the number of different elements in a sample*, Sib. Zh. Ind. Mat., **17**:3 (2014), 135–147 (in Russian).

[7] M.G. Chebunin, *Functional central limit theorem in an infinite urn scheme for distributions with superheavy tails*, Sib. Elektron. Mat. Izv. **14** (2017), 1289–1298.

[8] A. Pananjady, V. Muthukumar, A. Thangaraj, Just Wing It: Near-Optimal Estimation of Missing Mass in a Markovian Sequence, Journal of Machine Learning Research**25** ,(2024), 1–43

[9] Khmaladze, É. V, *Convergence properties in certain occupancy problems including the Karlin–Rouault law.*, Journal of Applied Probability, **48** No. 4, (2011), 1095—1113.

[10] R.R. Bahadur, *On the number of distinct values in a large sample from an infinite discrete distribution*, Proceedings of the National Institute of Sciences of India, **26A**, Supp. II (1960), 67–75.

[11] M. Chebunin, A. Kovalevskii, *Functional central limit theorems for certain statistics in an infinite urn scheme*, Statistics and Probability Letters, **119** (2016), 344–348.

[12] M. Chebunin, A. Kovalevskii, *Asymptotically normal estimators for Zipf's law*, Sankhya A (2019), **81**, 482–492.

[13] G. Decrouez, M. Grabchak, Q. Paris, *Finite sample properties of the mean occupancy counts and probabilities*, Bernoulli **24** (2018), no. 3, 1910–1941

[14] M. Dutko, *Central limit theorems for infinite urn models*, Ann. Probab. **17** (1989), 1255–1263.

[15] A. Gnedin, B. Hansen, J. Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws* , Probability Surveys **4** (2007), 146–171.

[16] E. S. Key, Rare Numbers. Journal of Theoretical Probability, Vol. 5, No. 2, (1992) 375–389.

[17] N.S. Zakrevskaya, A.P. Kovalevskii, One-parameter probabilistic models of text statistics , Sib. Zh. Ind. Mat. Vol. 4, No. 2,(2001) 142-153.

[18] O. Durieu G. Samorodnitsky Y. Wang, *From infinite urn schemes to self-similar stable processes*, Stochastic Processes and their Applications Volume 130, Issue 4, April 2020, Pages 2471-2487

[19] J. Garza, Y. Wang, A functional central limit theorem for weighted occupancy processes of the Karlin model, Stochastic Processes and their Applications Volume 188, October 2025, 104665

[20] S. Fayzullaev, A. Kovalevskii, *Hapax legomena via stochastic processes* , Glottometrics, Issue 56, (2024), pp. 22–39

[21] H. A. Simon, *On a class of skew distribution functions*, Biometrika, 42:3-4 (1955), 425-440.

[22] M. G. Chebunin, A. P. Kovalevskii, *Modifications of Karlin and Simon text models*, Siberian Electronic Mathematical Reports, vol. 19, (2022),p. 708-723.

Shahzod Shuxrat ugli Fayzullaev
Novosibirsk State University,
Pirogova st., 1
630090, Novosibirsk, Russia
Email address: s.faizullaev@g.nsu.ru

Artyom Pavlovich Kovalevskii
Sobolev Institute of Mathematics,
pr. Koptyuga, 4,
630090, Novosibirsk, Russia
Email address: artyom.kovalevskii@gmail.com