S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru

Том 19, №2, стр. 708–723 (2022) УДК 519.233
DOI 10.33048/semi.2022.19.059 MSC 62F03

MODIFICATIONS OF KARLIN AND SIMON TEXT MODELS

M.G. CHEBUNIN, A.P. KOVALEVSKII

Abstract. We discuss probability text models and their modifications. We construct processes of different and unique words in a text. The models are to correspond to the real text statistics. The infinite urn model (Karlin model) and the Simon model are the most known models of texts, but they do not give the ability to simulate the number of unique words correctly. The infinite urn model give sometimes the incorrect limit of the relative number of unique and different words. The Simon model states a linear growth of the numbers of different and unique words. We propose three modifications of the Karlin and Simon models. The first one is the offline variant, the Simon model starts after the completion of the infinite urn scheme. We prove limit theorems for this modification in embedded times only. The second modification involves repeated words in the Karlin model. We prove limit theorems for it. The third modification is the online variant, the Simon redistribution works at any toss of the Karlin model. In contrast to the compound Poisson model, we have no analytics for this modification. We test all the modifications by the simulation and have a good correspondence to the real texts.

Keywords: probability text models, Simon model, infinite urn model, weak convergence.

1. Introduction

Probabilistic text modeling involves several simplifications. However, the probabilistic model should maintain the behavior of text statistics that are observed

---

Chebunin, M.G., Kovalevskii, A.P., Modifications of Karlin and Simon text models.

The reported study was funded by RFBR and CNRS according to the research project No. 19-51-15001.

Received December, 26, 2021, published September, 7, 2022.

in practice. In particular, we will consider the number of different words in the text and the number of words that occur once.

Let R_n be a number of different words among the first n words of the text. $R_{n,i}$ be a number of words encountered i times, $R^*_{n,i}$ be the number of words encountered not lesser than i times. Therefore $R_n = R^*_{n,1}$, $R_{n,i} = R^*_{n,i} - R^*_{n,i+1}$, $i \geq 1$.

The power law of the growth of the number of different words is called Herdan's Law or Heaps' Law. It refers to Herdan [20] and Heaps [19].

Bahadur [4] and Karlin [23] studied an infinite urn model: any new ball goes to some of infinitely many urns with probability that corresponds to a power law and independently of anything else. The interpretation for texts corresponds words of a text to balls and words of a dictionary to urns.

Simon [27] proposed quite another model: the $(n+1)$-th word of a text is a new one with probability p, and coincides with any of previous words with probability $(1-p)/n$.

The infinite urn scheme looks more suitable for describing real texts, since Simon's model leads to a linear increase in the number of different words. However, the infinite urn scheme is not flexible enough to describe texts. We study two different estimates for the parameter θ of the exponential decay of the probabilities. One of them is $\widehat{\theta}$, it characterizes the rate at which the number of different words grows. Another estimate θ^* is the ratio of the number of unique words to the number of different words. According to the infinite urn scheme with exponential decay of probabilities, these two estimates should converge to the same number θ.

But we show in Section 2 the examples that the estimates are substantially different, the number of words encountered once (unique words) grows according to a power law with the same exponent but with a lower constant.

So we need some modifications or combinations of the models.

In Section 3, we study an elementary probabilistic text model. This is an infinite urn scheme. In this model, the number of different words and the number of words encountered once were studied by Bahadur[4], Karlin [23], Chebunin and Kovalevskii [12]. We study the correspondence between the empirical and theoretical behavior of these statistics.

In Section 4, we study Simon's model. Simon [27] proposed the next stochastic model: the $(n+1)$-th word in the text is new with probability p; it coincides with each of the previous words with probability $(1-p)/n$. In fact, he proposed a more general model with the same dynamics of numbers of word occurrences. He based his model on the model of Yule [28] who constructed it to explain the distribution of biological genera by number of species. Baur and Bertoin [9] proposed a modification of the Yule-Simon model with a wide class of limiting distributions.

We study the asymptotic behavior of the statistics $R_{n,1}$ in the Simon model based on functional limit theorems for urn models obtained by Janson.

In Section 5, we propose the offline Simon modification of the Karlin model. The purpose of these modifications is to correspond the theoretical and empirical behavior of the sequences $\{R_j\}_{j\geq 1}$ and $\{R_{j,1}\}_{j\geq 1}$. We prove analytical theorems (SLLN and FCLT) for the process in embedded times of increation of the initial urn process.

In Section 6, we study the second modification. It involves the compound Poisson process in the infinite urn model. We prove SLLN and FCLT for the modification. In Section 7, we propose the third modification. It is the online variant, the Simon redistribution works at any toss of the Karlin model. In contrast to the compound Poisson model, we have no limit theorems for this modification.

We test all the modifications by the simulation and have a good correspondence to the real texts. We discuss the advantages and disadvantages of the models in Section 8.

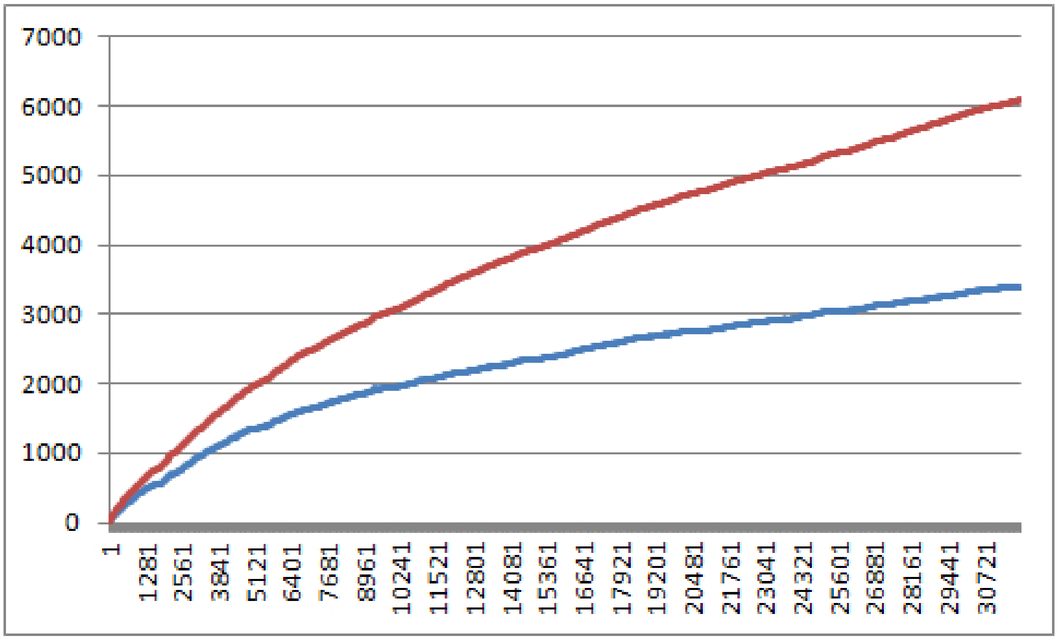

Fig. 1. Numbers of different words (red) and numbers of unique words (blue) in *Childe Harold's Pilgrimage* by Byron

2. Empirical analysis

We analyze the number of different words and the number of words encountered once in texts of different authors. These two processes behave like power functions with the same exponent but with different factors.

We estimate the exponent $\theta \in (0,1)$ of the power functions in two different ways. Chebunin and Kovalevskii [14] proposed the estimate

$$\widehat{\theta} = \log_2 R_n - \log_2 R_{[n/2]}$$

and studied conditions of its consistency and asymptotic normality. It has been used for the analysis of short texts (Zakrevskaya and Kovalevskii [29]).

Another estimate is

$$\theta^* = R_{n,1}/R_n.$$

Karlin [23] proved that it is consistent for the elementary text model under weak assumptions (see the next section). This is the asymptotically normal estimate under some additional assumptions (Chebunin and Kovalevskii [13]).

For *Childe Harold's Pilgrimage* by Byron we have $n = 37064$, $R_n = 6911$, $R_{n/2} = 4582$, $\widehat{\theta} = 0.5929$, $R_{n,1} = 3912$, $\theta^* = 0.5661$, so the second estimate is significantly smaller than the first one.

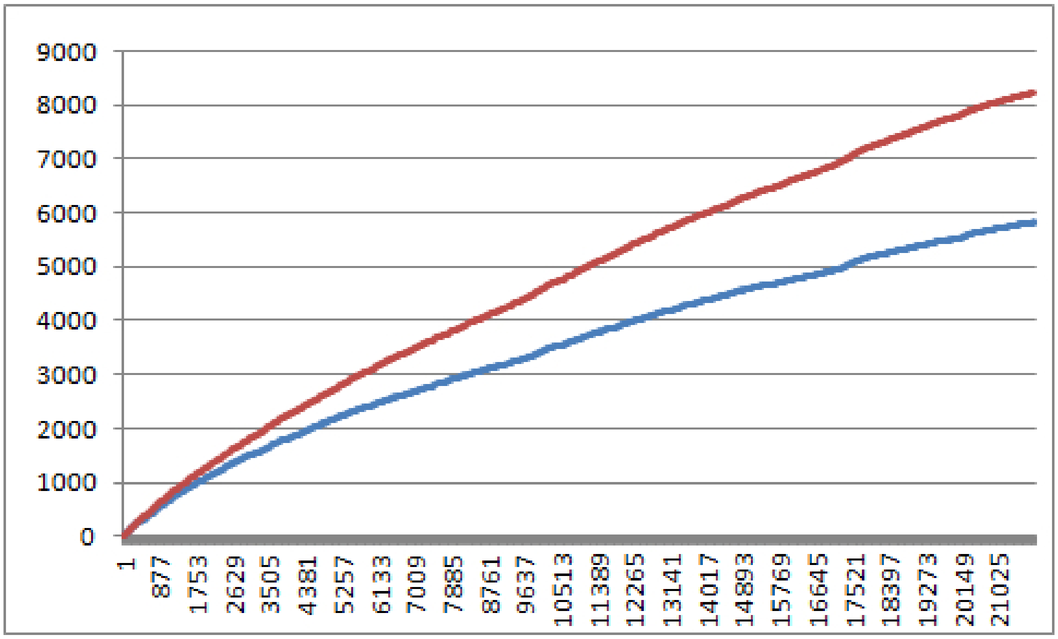

FIG. 2. Numbers of different words (red) and numbers of unique words (blue) in *Evgene Onegin* by Pushkin

For *Evgene Onegin* by Pushkin (in Russian) we have $n = 21882$, $R_n = 8236$, $R_{n/2} = 4916$, $\widehat{\theta} = 0.7445$, $R_{n,1} = 5824$, $\theta^* = 0.7071$, so the second estimate is significantly smaller than the first one.

3. Results for the elementary urn model

The simplest probabilistic model of text is the infinite urn scheme. Words are selected sequentially independently of each other from an infinite dictionary. The probabilities of the appearance of words decrease in accordance with the power distribution according to Zipf's law. As Bahadur showed, the number of different words is growing according to a power law. Karlin showed that the number of words met once grows under this model also according to a power law with the same exponent.

Karlin [23] studied an infinite urn scheme, that is, n balls distributed to urns independently and randomly; there are infinitely many urns. Each ball goes to urn i with probability $p_i > 0$, $p_1 + p_2 + \ldots = 1$ (without loss of generality $p_1 \geq p_2 \geq \ldots$).

Let (see Karlin [23]) $\Pi = \{\Pi(t),\ t \geq 0\}$ be a Poisson process with parameter 1. We denote by $X_i(n)$ a number of balls in urn i. According to well-known property of splitting of Poisson flows, stochastic processes $\{X_i(\Pi(t)),\ t \geq 0\}$ are Poisson with intensities p_i and are mutually independent for different i's. The definition implies that

$$R^*_{\Pi(t),k} = \sum_{i=1}^{\infty} \mathbf{I}(X_i(\Pi(t)) \geq k), \quad R_{\Pi(t),k} = \sum_{i=1}^{\infty} \mathbf{I}(X_i(\Pi(t)) = k).$$

Let designate $\alpha(x) = \max\{j | \ p_j \geq 1/x\}$.

We assume, following Karlin [23], that

$$\alpha(x) = x^\theta L(x), \quad 0 < \theta < 1.$$

Here $L(x)$ is a slowly varying function as $x \to \infty$. Let for $t \in [0,1], \ k \geq 1$

$$Y^*_{n,k}(t) = \frac{R^*_{[nt],k} - \mathbf{E}R^*_{[nt],k}}{(\alpha(n))^{1/2}}, \qquad Z^*_{n,k}(t) = \frac{R^*_{\Pi(nt),k} - \mathbf{E}R^*_{\Pi(nt),k}}{(\alpha(n))^{1/2}},$$

$$Y_{n,k}(t) = \frac{R_{[nt],k} - \mathbf{E}R_{[nt],k}}{(\alpha(n))^{1/2}}, \qquad Z^{**}_{n,k}(t) = \frac{R^*_{\Pi([nt]),k} - \mathbf{E}R^*_{\Pi([nt]),k}}{(\alpha(n))^{1/2}},$$

$K_{k,\theta} = \theta\Gamma(k-\theta)$ for $k > 0$ and $K_{k,\theta} = -\Gamma(1-\theta)$ for $k = 0$.

Proposition 1. *(Theorem 4 in [23]). Let $\theta \in (0,1]$. Then $(R_n - \mathbf{E}R_n)/B_n^{1/2}$ converges weakly to standard normal distribution, where*

$$B_n = \begin{cases} \Gamma(1-\theta)(2^\theta - 1)n^\theta L(n), & \theta \in (0,1); \\ n\int\limits_0^\infty \frac{e^{-1/y}}{y} L(ny)dy \overset{def}{=} nL^*(n), & \theta = 1; \end{cases}$$

Karlin ([23], Lemma 4) proved function $L^*(x)$ to be slowly varying as $x \to \infty$.

Dutko [16] generalized the theorem by proving asymptotic normality of R_n if $\mathbf{Var}R_n \to \infty$ as $n \to \infty$. This condition always holds if $\theta \in (0,1]$ but can hold too for $\theta = 0$. Gnedin, Hansen and Pitman [17] focus on the study of conditions for convergence $\mathbf{Var}R_n \to \infty$. They also collect facts on the issue.

Proposition 2. *(Theorem 5 in [23]). Let $\theta \in (0,1)$, $r_1 < \ldots < r_\nu$ be ν positive integers. Then random vector $(Y_{n,r_1}(1), \ldots, Y_{n,r_\nu}(1))$ converges weakly to a multivariate normal distribution with zero expectation and covariances*

$$c_{r_i,r_j} = \begin{cases} -\frac{\theta\Gamma(r_i+r_j-\theta)}{r_i! r_j!} 2^{\theta-r_i-r_j}, & i \neq j; \\ \frac{\theta}{\Gamma(r_i+1)} \left(\Gamma(r_i-\theta) - 2^{-2r_i+\theta} \frac{\Gamma(2r_i-\theta)}{\Gamma(r_i+1)}\right), & i = j. \end{cases}$$

Barbour and Gnedin [6] extended this result to the case $\theta = 0$ if variances go to infinity. They found conditions for convergence of covariances to a limit and identified four types of limiting behavior of variances. Barbour [5] proved theorems on the approximation of the number of cells with k balls by translated Poisson distribution. Key [24], [25] studied the limit behavior of statistics $R_{n,1}$. Hwang and Janson [21] proved local limit theorems for finite and infinite numbers of cells. Chebunin [11] constructed R_n-based explicit parameter estimators for a wide range of one-parameter families and proved their consistency.

Durieu and Wang [15] established a functional central limit theorem for randomization of a process R_n: indicators are multiplied randomly by ± 1 before summing. The limiting Gaussian process is a sum of independent self-similar processes in this case.

Chebunin & Kovalevskii [12] proved the next FCLT.

Proposition 3. **(i)** *Let* $\theta \in (0,1)$, $\nu \geq 1$ *is integer. Then process* $\left(Y^*_{n,1}(t),\dots,Y^*_{n,\nu}(t),\ 0\le t\le 1\right)$ *converges weakly in the uniform metrics in* $D\left([0,1]^\nu\right)$ *to* ν*-dimensional Gaussian process with zero expectation and covariance function* $(c^*_{ij}(\tau,t))^{\nu}_{i,j=1}$*: for* $\tau\le t$, $i,j\in\{1,\dots,\nu\}$ *(taking* $0^0=1$*)*

$$c^*_{ij}(\tau,t)=\begin{cases}\sum_{s=0}^{i-1}\sum_{m=0}^{j-s-1}\frac{\tau^s(t-\tau)^m K_{m+s,\theta}}{t^{m+s-\theta}s!m!}-\sum_{s=0}^{i-1}\sum_{m=0}^{j-1}\frac{\tau^s t^m K_{m+s,\theta}}{(t+\tau)^{m+s-\theta}s!m!}, & i<j;\\ t^\theta\sum_{m=0}^{j-1}\frac{K_{m,\theta}}{m!}-\sum_{s=0}^{i-1}\sum_{m=0}^{j-1}\frac{\tau^s t^m K_{m+s,\theta}}{(t+\tau)^{m+s-\theta}s!m!}, & i\ge j;\end{cases}$$

$c^*_{ij}(\tau,t)=c^*_{ji}(t,\tau)$.

(ii) *Let* $\theta=1$*. Then process* $\left(\frac{R_{[nt]}-\mathbf{E}R_{[nt]}}{(nL^*(n))^{1/2}},\ 0\le t\le 1\right)$ *converges weakly in the uniform metrics in* $D([0,1])$ *to a standard Wiener process.*

4. Results for Simon model

Yule showed that, for the Yule—Simon model,

$$\mathbf{E}R_{n,i}/\mathbf{E}R_n\to f(i),\quad i\ge 1, \tag{1}$$

$$f(i)=\rho B(i,1+\rho),$$

$\rho=(1-p)^{-1}$, $B(\cdot,\cdot)$ is Beta function.

We analyze stochastic aspects of this convergence. The limiting distribution is named Yule-Simon distribution.

There are many ramifications and applications of the Yule-Simon model. Haight & Jones [18] gave special references to word associations tests. Lansky & Radill-Weiss [26] proposed a generalization of the model for better correspondence to applications.

This model can be embedded in a more general context of random cutting of recursive trees. In this context, statistics under study are the most frequent words. See Aldous & Pitman [3] for its limiting distribution and convergence, Baur & Bertoin [7], [8] for an overview and new results.

Aldous [2] proposed a generalization of the limiting distribution but without an underlying process.

Janson [22] considered generalized Polya urns and proved SLLN, CLT and FCLT for it. Finite-dimensional vectors

$$(R_n,R_{n,1},\dots,R_{n,m-1},n-\sum_{i<m}iR_{n,i})$$

can be studied using these models. So we have componentwise SLLN and finite-dimensional CLT and FCLT for these statistics.

Theorem 1. *For any* $p\in(0,1)$ *in Simon model and any* $m>1$

$$\frac{R_n}{n}\to p\quad\text{a.s.},$$

$$\frac{(R_{n,1},\dots,R_{n,m-1})}{n}\to\frac{p}{1-p}\left(B\left(1,\frac{2-p}{1-p}\right),\dots,B\left(m-1,\frac{2-p}{1-p}\right)\right)\quad\text{a.s.},$$

and, in $D[0,\infty)$,

$$\Big\{n^{-1/2}(R_{[nt]}-tnp),$$

$$n^{-1/2}\left(R_{[nt],j}-tn\frac{p}{1-p}B\left(j,\frac{2-p}{1-p}\right)\right),\ 1\le j\le m-1,\ t\ge 0\Bigg\}\to_d V(t),$$

where $V(t)$ is a continuous centered m-dimensional Gaussian process, its covariance matrix-function $\mathbf{E}V(x)V^T(y)$ depends on p,x,y only.

Proof

Simon's model can be studied as a very partial case of Janson's [22] urn scheme. In this model, Simon urns with $1\le i\le m-1$ balls are balls with numbers from 2 to m with weights $a_i=i-1$. The $(m+1)$-th urn contains all other balls with weights $a_{m+1}=1$. Balls with number 1 correspond to all different words, $a_1=0$. So the random uniform choice of balls in the Simon model corresponds to the random choice with weights a_i, $1\le i\le m+1$, in Janson model.

Let $(\mathbf{e}_1,\dots,\mathbf{e}_{m+1})$ be the standard basis in $\mathbf{R}^{m+1}$, then increment vectors in Janson model are

$$\xi_i=\mathbf{e}_1+\mathbf{e}_2 \text{ with probability } p,\ \ \xi_i=\mathbf{e}_{i+1}-\mathbf{e}_i \text{ with probability } q=1-p,\ \ 2\le i<m,$$

$$\xi_m=\mathbf{e}_1+\mathbf{e}_2 \text{ with probability } p,\quad \xi_m=m\mathbf{e}_{m+1}-\mathbf{e}_m \text{ with probability } q=1-p,$$

$$\xi_{m+1}=\mathbf{e}_1+\mathbf{e}_2 \text{ with probability } p,\quad \xi_{m+1}=\mathbf{e}_{m+1} \text{ with probability } q=1-p.$$

So matrix $A=(a_j\mathbf{E}\xi_{ji})$ has the first (maximal) eigenvalue $\lambda_1=1$, the next (second) eigenvalue $\lambda_2=0$. Thus the assumption $\lambda_2<\lambda_1/2$ of Theorem 2.22 and Theorem 2.31(i) in Janson [22] is hold. So we use Janson's theorems 2.21 (SLLN), 2.22, 2.31(i).

The proof is complete.

5. Offline Simon modification of the Karlin model

Let the balls take urns as in the Karlin model, but the appearance of a new non-empty urn corresponds only to the embryo of a new word.

After the finish of the Karlin model, Simon's model starts to work in moments of the appearance of the embryos of new words: the k-th embryo coincides with each of the previous embryos with probability q/k, and is a new word with probability p, $p+q=1$. We use Janson's theorems.

Balls of the first type are embryos of words encountered exactly once, their number at the k -th step is $X_{k,1}$, balls of this type have weights (responsible for the probabilities of choosing balls) $a_1=1$.

Balls of the second type are embryos of words encountered more than once, their number at the k -th step is $X_{k,2}$, balls of this type have weights $a_2=1$ also.

The numbers $X_{k,1}$ and $X_{k,2}$ form vectors

$$X_k=(X_{k,1},X_{k,2})',\quad k\ge 0.$$

The system starts from the state $X_{0,1}=1,X_{0,2}=0$. At each step k, one ball is chosen at random from $X_{k,1}+X_{k,2}$ balls of the first and second types.

If the selected ball is of the first type, then a ball of the first type is added with probability p, and it is replaced by a ball of the second type with probability q.

The evolution of vector X_k is done by Janson [22] as follows: "The drawn ball is returned to the urn together with ΔX_{kj} balls of type j, for each $j=1,2$, where $\Delta X_k=(\Delta X_{k,1},\Delta X_{k,2})'$ is a random vector such that if the drawn ball has type i, then ΔX_k has the same distribution as ξ_i and is independent of everything else that has happened so far". We have

$$X_0=(1,0)',$$

$$\xi_1 = (1,0)' \text{ with probability } p,$$
$$\xi_1 = (-1,1)' \text{ with probability } q.$$

If the chosen ball is of the second type, then a ball of the first type is added with probability p, and nothing happens with probability q:

$$\xi_2 = (1,0)' \text{ with probability } p,$$
$$\xi_2 = (0,0)' \text{ with probability } q.$$

So

$$A = (a_j \mathbf{E}\xi_{ji})_{i,j=1}^2 = \begin{pmatrix} p-q & p \\ q & 0 \end{pmatrix}.$$

Eigenvalues of A are $\lambda_1 = p$, $\lambda_2 = -q < \lambda_1/2$.

Eigenvector v_1 of A correspond to eigenvalue $\lambda_1 = p$ and condition (2.2) from Janson [22] $a \cdot v_1 = 1$. As $a = (1\,1)'$, we have

$$v_1 = \begin{pmatrix} p \\ q \end{pmatrix}.$$

We calculate e^{tA} using Sylvester's formula:

$$e^{tA} = \frac{e^{pt}}{p+q}(A+qI) - \frac{e^{-qt}}{p+q}(A-pI) = e^{pt}(A+qI) - e^{-qt}(A-pI).$$

From (3.17) in Janson [22],

$$\phi(s,A) = \int_0^s e^{tA}dt = \frac{e^{ps}-1}{p}(A+qI) + \frac{e^{-qs}-1}{q}(A-pI)$$

From (3.18) in [22],

$$\psi(s,A) = e^{sA} - \lambda_1 v_1 a' \phi(s,A)$$
$$= e^{ps}(A+qI) - e^{-qs}(A-pI) - pv_1a'\left(\frac{e^{ps}-1}{p}(A+qI) + \frac{e^{-qs}-1}{q}(A-pI)\right).$$

Note that

$$v_1 a' = A + qI = (A+qI)^2.$$

So

$$\psi(s,A) = -e^{-qs}(A-pI) + A + qI - \frac{p(e^{-qs}-1)}{q}(A+qI)(A-pI).$$
$$= A + qI - e^{-qs}(A-pI).$$

Let

$$A_0 := A + qI,$$

then

$$\psi(s,A) = A_0(1-e^{-qs}) + Ie^{-qs}.$$

We calculate matrices B_1, B_2, B. From (2.13) in [22],

$$B_1 = \mathbf{E}\xi_1\xi_1' = p\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + q\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -q \\ -q & q \end{pmatrix},$$

$$B_2 = \mathbf{E}\xi_2\xi_2' = \begin{pmatrix} p & 0 \\ 0 & 0 \end{pmatrix}.$$

From (2.14) in [22],

$$B = a_1 v_{11} B_1 + a_2 v_{12} B_2 = p\begin{pmatrix} 1+q & -q \\ -q & q \end{pmatrix}.$$

From Theorem 3.21 in [22],

$$\frac{X_n}{n} \to_{a.s.} \lambda_1 v_1 = p v_1.$$

From Theorem 3.22 in [22],

$$\frac{X_n - n\lambda_1 v_1}{\sqrt{n}} \to_d N(0,\Sigma),$$

$$\Sigma = \int_0^\infty \psi(s,A)B\psi(s,A)' e^{-\lambda_1 s}\lambda_1 ds - \lambda_1^2 v_1 v_1'$$

$$= \int_0^\infty \left(A_0(1-e^{-qs}) + Ie^{-qs}\right) B \left(A_0'(1-e^{-qs}) + Ie^{-qs}\right) e^{-ps} p ds$$

$$-p^2 \begin{pmatrix} p^2 & pq \\ pq & q^2 \end{pmatrix}$$

$$= \int_0^\infty \left(A_0BA_0' + e^{-qs}(A_0B + BA_0' - 2A_0BA_0') + e^{-2qs}(B + A_0BA_0')\right) e^{-ps} p ds$$

$$-p^2 \begin{pmatrix} p^2 & pq \\ pq & q^2 \end{pmatrix}$$

$$= A_0BA_0' + p(A_0B + BA_0' - 2A_0BA_0') + \frac{p}{1+q}(B + A_0BA_0') - p^2 \begin{pmatrix} p^2 & pq \\ pq & q^2 \end{pmatrix}$$

$$= pq \begin{pmatrix} 3p^2+p & p-3p^2 \\ p-3p^2 & 1-3p+3p^2 \end{pmatrix}.$$

We calculate using Theorem 3.31 in [22]:

$$\mathbf{E}U(x)U'(y) = \int\limits_{-\lambda_1^{-1}\log x}^{\infty} \psi(s+\lambda_1^{-1}\log x, A)B\psi(s+\lambda_1^{-1}\log y, A)' e^{-\lambda_1 x}\lambda_1 dx - x\lambda_1^2 v_1 v_1'$$

$$= \int\limits_{-\lambda_1^{-1}\log x}^{\infty} \left(A_0(1-e^{-q(s+\log x/p)}) + Ie^{-q(s+\log x/p)}\right) B$$

$$\times \left(A_0'(1-e^{-q(s+\log y/p)}) + Ie^{-q(s+\log y/p)}\right) e^{-\lambda_1 x}\lambda_1 dx - x\lambda_1^2 v_1 v_1'$$

$$= xpq \begin{pmatrix} 2p^2 & 2pq \\ pq-p^2 & q^2-pq \end{pmatrix} + xp^2q\left(\frac{x}{y}\right)^{q/p} \begin{pmatrix} 1+p & -1-p \\ -p & p \end{pmatrix}.$$

From Theorem 3.31 in [22], there is FCLT for $\{X_{[nt]},\ 0 \le t \le 1\}$:

Theorem 2. *There is convergence*

$$\left\{\frac{X_{[nt]} - nt\lambda_1 v_1}{\sqrt{n}},\ 0 \le t \le 1\right\} \to_d U = \{U(t),\ 0 \le t \le 1\}$$

in the uniform metrics, U is the centered Gaussian process, and for $0 < x \le y$

$$\mathbf{E}U(x)U'(y) = xpq \begin{pmatrix} 2p^2 & 2pq \\ pq-p^2 & q^2-pq \end{pmatrix} + xp^2q\left(\frac{x}{y}\right)^{q/p} \begin{pmatrix} 1+p & -1-p \\ -p & p \end{pmatrix}.$$

We have a good match for *Childe Harold's Pilgrimage* with $\theta = 0.58$, $p = 0.56$, for *Evgene Onegin* with $\theta = 0.74$, $p = 0.71$.

6. Modification of the Karlin model with repeated words

We want to invent a text model such that θ^* converges to a number less than $\widehat{\theta}$. We propose the following simple model: words appear in the same way as in Karlin's model, but each word with probability q_i (independently of others and of the process of the appearance of words) is repeated $i \geq 1$ times, $q_1 + q_2 + \ldots = 1$. New statistics of the number of different words and the number of words that occur once are denoted by $\mathbb{R}_n$ and $\mathbb{R}_{n,1}$. We denote by $\mathbb{X}_i(n)$ a number of balls in urn i, and

$$\mathbb{R}^*_{n,k} = \sum_{i=1}^{\infty} \mathbf{I}\left(\mathbb{X}_i(n) \geq k\right), \quad \mathbb{R}_n = \mathbb{R}^*_{n,1}, \quad \mathbb{R}_{n,k} = \mathbb{R}^*_{n,k} - \mathbb{R}^*_{n,k+1}.$$

The Poisson version of these processes uses the compound Poisson process. Let $\Pi = \{\Pi(t), t \geq 0\}$ be a Poisson process with parameter 1. The Poissonized version of Karlin model assumes the total number of $\Pi(n)$ balls. According to well-known thinning property of Poisson flows, stochastic processes $\left\{\mathbb{X}_i(\Pi(t)) \stackrel{\text{def}}{=} \mathbf{\Pi}_i(t), t \geq 0\right\}$ are compound Poisson with intensities p_i and are mutually independent for different i 's. The definition implies that

$$\mathbb{R}^*_{\Pi(n),k} = \sum_{i=1}^{\infty} \mathbf{I}\left(\mathbf{\Pi}_i(n) \geq k\right), \quad \mathbb{R}_{\Pi(n),k} = \sum_{i=1}^{\infty} \mathbf{I}\left(\mathbf{\Pi}_i(n) = k\right).$$

We proved FCLT for the vector process $(\mathbb{R}_n, \ \mathbb{R}_{n,1})$. Let $\alpha(x) = \max\{j \mid p_j \geq 1/x\}$. Following Karlin [23], we assume that $\alpha(x) = x^{\theta}L(x), 0 \leq \theta \leq 1$. Here $L(x)$ is a slowly varying function as $x \to \infty$. Let for $t \in [0,1], k \geq 1$

$$\mathbb{Y}^*_{n,k}(t) = \frac{\mathbb{R}^*_{[nt],k} - \mathbf{E}\mathbb{R}^*_{[nt],k}}{(\alpha(n))^{1/2}}, \quad \mathbb{Z}^*_{n,k}(t) = \frac{\mathbb{R}^*_{\Pi(nt),k} - \mathbf{E}\mathbb{R}^*_{\Pi(nt),k}}{(\alpha(n))^{1/2}},$$

$$\mathbb{Y}_{n,k}(t) = \frac{\mathbb{R}_{[nt],k} - \mathbf{E}\mathbb{R}_{[nt],k}}{(\alpha(n))^{1/2}}.$$

Theorem 3. *Let $\theta \in (0,1)$ and $v \geq 1$ be an integer. Then process $\left(\mathbb{Y}^*_{n,1}(t), \ldots, \mathbb{Y}^*_{n,v}(t), 0 \leq t \leq 1\right)$ converges weakly in the uniform metric in $D\left([0,1]^v\right)$ to v -dimensional Gaussian process with zero expectation and covariance function $\left(\mathbb{C}^*_{ij}(\tau,t)\right)_{i,j=1}^{v}$.*

Lemma 1. **(i)** *There exist $n_0 \geq 1$, $C(\theta) < \infty$ such that $\mathbf{E}R_{\Pi(n\delta)}/\alpha(n) \leq C(\theta)\delta^{\theta/2}$ for any $\delta \in [0,1], \ n \geq n_0$.*

(ii) *Let $\tau \leq t$, then $\mathbf{E}(\mathbb{R}^*_{\Pi(t),k} - \mathbb{R}^*_{\Pi(\tau),k}) \leq \mathbf{E}R_{\Pi(t-\tau)}, \ k \geq 1$.*

(iii) *For any pair ε, $\delta \in (0,1)$ there exists integer n_0 such that $\mathbf{P}(\forall t \in [0,1] \ \ \exists \tau : |\tau - t| \leq \delta, \ \Pi(n\tau) = [nt]) \stackrel{def}{=} \mathbf{P}(A(n)) \geq 1 - \varepsilon/2$ for any $n \geq n_0$.*

Proof. **(i)** and **(iii)** were proved in Chebunin and Kovalevskii [14], **(ii)** follows from

$$\mathbf{E}\left(\mathbb{R}^*_{\Pi(t),k} - \mathbb{R}^*_{\Pi(\tau),k}\right) = \sum_{i=1}^{\infty}\sum_{j=0}^{k-1} \mathbf{P}\left(\mathbf{\Pi}_i(\tau) = j\right) \mathbf{P}\left(\mathbf{\Pi}_i(t) - \mathbf{\Pi}_i(\tau) \geq k - j\right)$$

$$\leq \sum_{i=1}^{\infty} \mathbf{P}\left(\mathbf{\Pi}_i(t-\tau) \geq 1\right) = \mathbf{E}\mathbb{R}_{\Pi(t-\tau)} = \mathbf{E}R_{\Pi(t-\tau)}.$$

The proof is complete.

Proof of Theorem 3

Step 1 (covariances) Let $\tau \le t$, $i \le j$ and $\pi_{k,i}(t) = \mathbf{P}(\mathbf{\Pi}_k(t) = i)$

$$\mathbf{cov}\left(\mathbb{R}^*_{\Pi(\tau),i}, \mathbb{R}^*_{\Pi(t),j}\right) = \sum_{k=1}^{\infty} \left(\mathbf{P}\left(\mathbf{\Pi}_k(\tau) < i, \mathbf{\Pi}_k(t) < j\right) - \mathbf{P}\left(\mathbf{\Pi}_k(\tau) < i\right)\mathbf{P}\left(\mathbf{\Pi}_k(t) < j\right)\right)$$

$$= \sum_{k=1}^{\infty}\sum_{s=0}^{i-1} \pi_{k,s}(\tau)\left(\sum_{m=0}^{j-s-1}\pi_{k,m}(t-\tau) - \sum_{m=0}^{j-1}\pi_{k,m}(t)\right),$$

and $i > j$

$$\mathbf{cov}\left(\mathbb{R}^*_{\Pi(\tau),i}, \mathbb{R}^*_{\Pi(t),j}\right) = \sum_{k=1}^{\infty}\sum_{s=0}^{j-1} \pi_{k,s}(t)\left(1 - \sum_{m=0}^{i-1}\pi_{k,m}(\tau)\right).$$

Let $t' = t(1-q_0)$, $c_1 = q_1/(1-q_0)$

$$\mathbf{E}\mathbb{R}_{\Pi(t)} = \sum_{k=1}^{\infty} 1 - \pi_{k,0}(t) = \sum_{k=1}^{\infty} 1 - e^{-p_k t} = \mathbf{E}R_{\Pi(t)}, \quad \mathbf{D}\mathbb{R}_{\Pi(t)} = \mathbf{D}R_{\Pi(t)}.$$

$$\mathbf{E}\mathbb{R}_{\Pi(t),1} = \sum_{k=1}^{\infty} \pi_{k,1}(t) = \sum_{k=1}^{\infty} q_1 \mathbf{P}(\Pi_k(t) = 1) = q_1 \mathbf{E}R_{\Pi(t),1}.$$

$$\mathbf{cov}\left(\mathbb{R}^*_{\Pi(\tau)}, \mathbb{R}^*_{\Pi(t)}\right) = \sum_{k=1}^{\infty} \pi_{i,0}(t)(1-\pi_{i,0}(\tau)) = \mathbf{E}R_{\Pi(t+\tau)} - \mathbf{E}R_{\Pi(t)}.$$

$$\mathbf{cov}\left(\mathbb{R}^*_{\Pi(\tau),2}, \mathbb{R}^*_{\Pi(t),2}\right) = \sum_{k=1}^{\infty}\sum_{s=0}^{1} \pi_{k,s}(\tau)\left(\sum_{m=0}^{1-s}\pi_{k,m}(t-\tau) - \sum_{m=0}^{1}\pi_{k,m}(t)\right)$$

$$= \sum_{k=1}^{\infty} \pi_{i,0}(t)(1-\pi_{i,0}(\tau)) + e^{-p_k\tau}(q_1 p_k (t-\tau) e^{-p_k(t-\tau)} - q_1 p_k t e^{-p_k t})$$

$$+ \sum_{k=1}^{\infty} q_1 p_k \tau e^{-p_k \tau}(e^{-p_k(t-\tau)} - e^{-p_k t} - q_1 p_k t e^{-p_k t})$$

$$= \mathbf{E}(R_{\Pi(t+\tau)} - R_{\Pi(t)}) + q_1 \mathbf{E}R_{\Pi(t),1} - q_1 \mathbf{E}R_{\Pi(t+\tau),1} - \frac{2q_1^2 t\tau}{(t+\tau)^2}\mathbf{E}R_{\Pi(t+\tau),2}.$$

$$\mathbf{cov}\left(\mathbb{R}^*_{\Pi(\tau)}, \mathbb{R}^*_{\Pi(t),2}\right) = \sum_{k=1}^{\infty} \pi_{k,0}(\tau)\left(\sum_{m=0}^{1}\pi_{k,m}(t-\tau) - \sum_{m=0}^{1}\pi_{k,m}(t)\right)$$

$$= \mathbf{E}(R_{\Pi(t+\tau)} - R_{\Pi(t)}) + \frac{q_1(t-\tau)}{t}\mathbf{E}R_{\Pi(t),1} - \frac{q_1 t}{t+\tau}\mathbf{E}R_{\Pi(t+\tau),1}.$$

$$\mathbf{cov}\left(\mathbb{R}^*_{\Pi(t)}, \mathbb{R}^*_{\Pi(\tau),2}\right) = \sum_{k=1}^{\infty} \pi_{k,0}(t)\left(1 - \pi_{k,0}(\tau) - \pi_{k,1}(\tau)\right)$$

$$= \mathbf{E}(R_{\Pi(t+\tau)} - R_{\Pi(t)}) - \frac{q_1 \tau}{t+\tau}\mathbf{E}R_{\Pi(t+\tau),1}.$$

Since

$$\mathbf{E}R_{\Pi(t)} \sim \Gamma(1-\theta)\alpha(t), \ \mathbf{E}R_{\Pi(t),k} \sim \theta\frac{\Gamma(k-\theta)}{k!}\alpha(t) \ \text{ if } \ \theta \in (0,1), \ k \ge 1,$$

then

$$\mathbb{C}^*_{11}(\tau, t) = \Gamma(1-\theta)((t+\tau)^\theta - t^\theta),$$

$$\mathbb{C}_{22}^*(\tau,t)=\Gamma(1-\theta)\left(((t+\tau)^\theta-t^\theta)(1-q_1\theta)-\frac{q_1^2\theta(1-\theta)t\tau}{(t+\tau)^{2-\theta}}\right),$$

$$\mathbb{C}_{12}^*(\tau,t)=\Gamma(1-\theta)\left((t+\tau)^\theta-t^\theta+\frac{q_1\theta(t-\tau)}{t^{1-\theta}}-\frac{q_1\theta t}{(t+\tau)^{1-\theta}}\right),$$

$$\mathbb{C}_{12}^*(t,\tau)=\Gamma(1-\theta)\left((t+\tau)^\theta-t^\theta-\frac{q_1\theta\tau}{(t+\tau)^{1-\theta}}\right).$$

Step 2 (convergence of finite-dimensional distributions) Analogously to proof of Theorem 1 in [16], we have that, for any fixed $m \geq 1, 0 < t_1 < t_2 < \cdots < t_m \leq 1$ triangle array of mv -dimensional random vectors $\left\{((\mathbf{I}(\mathbf{\Pi}_k(nt_j)\geq i)-\mathbf{P}(\mathbf{\Pi}_k(nt_j)\geq i))\alpha^{-1/2}(n), i\leq v, j\leq m), k\leq n\right\}_{n\geq 1}$ satisfies Lindeberg condition.

Step 3 (relative compactness) Let for any $\tau_1 \leq \tau_2$,

$$\mathbb{R}^*_{\Pi(\tau_2),k}-\mathbb{R}^*_{\Pi(\tau_1),k}=\sum_{i=1}^\infty \mathbf{I}\left(\mathbf{\Pi}_i\left(\tau_2\right)\geq k,\mathbf{\Pi}_i\left(\tau_1\right)<k\right)\overset{\text{def}}{=}\sum_{i=1}^\infty \mathbb{I}_i\left(\tau_1,\tau_2\right)=\sum_{i=1}^\infty \mathbb{I}_i,$$

$\mathbb{P}_i=\mathbb{P}_i\left(\tau_1,\tau_2\right)=\mathbf{P}\left(\mathbb{I}_i\right)$. We will use designations $\mathbb{I}_i$ and corresponding $\mathbb{P}_i$ for different values of $\tau_1<\tau_2$.

We need a new process $\mathbb{Z}^{**}_{n,k}(t)=\frac{\mathbb{R}^*_{\Pi([nt]),k}-\mathbf{E}\mathbb{R}^*_{\Pi([nt]),k}}{(\alpha(n))^{1/2}}$.

We (a) prove continuity of the limiting process; (b) prove that $\mathbb{Z}^*_{n,k}$ and $\mathbb{Z}^{**}_{n,k}$ are 'close'; (c) prove relative compactness of $\mathbb{Z}^{**}_{n,k}$.

(a) Let $\tau_1=nt_1, \tau_2=nt_2$ for $t_1<t_2$, then

$$\mathbf{E}\left(\mathbb{Z}^*_{n,k}\left(t_2\right)-\mathbb{Z}^*_{n,k}\left(t_1\right)\right)^2=\sum_{i=1}^\infty \mathbf{E}\left(\mathbb{I}_i-\mathbb{P}_i\right)^2/\alpha(n)$$

$$\leq\sum_{i=1}^\infty \mathbb{P}_i/\alpha(n)\leq C(\theta)\left(t_2-t_1\right)^{\theta/2}.$$

Here we used the fact that variance of an indicator is lesser than its expectation and Lemma 1 (i, ii). Using Step 1 and Theorem 1.4 in Adler [1], we prove that the k-th component of the limiting Gaussian process is in $C(0,1)$ a.s. As the limiting Gaussian process is in $C\left([0,1]^v\right)$ a.s., weak convergence in Skorokhod topology implies the same in the uniform topology.

(b) As $\mathbb{R}^*_{\Pi(nt),k}-\mathbb{R}^*_{\Pi([nt]),k}\leq \Pi([nt]+1)-\Pi([nt])$ a.s., and $\mathbf{E}(\Pi([nt]+1)-\Pi([nt]))=1$ we have for any $\eta>0$

$$\mathbf{P}\left(\sup_{0\leq t\leq 1}\left|\mathbb{Z}^*_{n,k}(t)-\mathbb{Z}^{**}_{n,k}(t)\right|>\eta\right)$$

$$\leq\mathbf{P}\left(\sup_{0\leq m\leq n}(\Pi(m+1)-\Pi(m)+1)>\eta\sqrt{\alpha(n)}\right)$$

$$\leq\sum_{m=0}^n \mathbf{E}e^{\Pi(m+1)-\Pi(m)+1}/e^{\eta\sqrt{\alpha(n)}}=(n+1)e^{e-\eta\sqrt{\alpha(n)}}\to 0$$

as $n\to\infty$. So it is enough to prove relative compactness of $\{\mathbb{Z}^{**}_{n.k}\}_{n\geq n_0}$.

(c) Let $t_2 - t_1 \geq \frac{1}{2n}$, then $[nt_2] - [nt_1] \leq n(t_2 - t_1) + 1 \leq 3n(t_2 - t_1)$. Let $\gamma = [16/\theta] + 1$, and $\tau_1 = [nt_1], \tau_2 = [nt_2]$. Using independence of terms and Rosenthal inequality, we have for all $n \geq n_0$ (where n_0 is from Lemma 1 (i))

$$\mathbf{E}\left|\mathbb{Z}^{**}_{n,k}(t_2) - \mathbb{Z}^{**}_{n,k}(t_1)\right|^\gamma \leq \frac{c(\gamma)}{(\alpha(n))^{\gamma/2}}\left(\sum_{i=1}^\infty \mathbf{E}\left|\mathbb{I}_i - \mathbb{P}_i\right|^\gamma + \left(\sum_{i=1}^\infty \mathbf{E}\left(\mathbb{I}_i - \mathbb{P}_i\right)^2\right)^{\gamma/2}\right)$$

$$\leq \frac{c(\gamma)}{(\alpha(n))^{\gamma/2}}\left(\sum_{i=1}^\infty \mathbb{P}_i + \left(\sum_{i=1}^\infty \mathbb{P}_i\right)^{\gamma/2}\right)$$

$$\leq \frac{c(\gamma)}{(\alpha(n))^{\gamma/2}}\left(24n^4(t_2 - t_1)^4 + \left(\mathbf{E}R_{\Pi(3n(t_2-t_1))}\right)^{\gamma/2}\right) \leq \widetilde{C}(\theta)(t_2 - t_1)^4.$$

Here $c(\gamma)$ and $\widetilde{C}(\theta)$ depend on its argument only. Above we used the fact that variance of an indicator is lesser than its expectation, inequality $\sum_i \mathbb{P}_i \leq \mathbf{E}(\Pi([nt_2]) - \Pi([nt_1])) \leq 3n(t_2 - t_1) \leq 24n^4(t_2 - t_1)^4$, and Lemma 1(i, ii).

Let $0 \leq t_2 - t_1 < 1/n$, then $[nt_1] = [nt]$ or $[nt_2] = [nt]$ for any $t \in [t_1, t_2]$. So $\mathbb{Q} \overset{\text{def}}{=} \mathbf{E}\left(\left|\mathbb{Z}^{**}_{n,k}(t) - \mathbb{Z}^{**}_{n,k}(t_1)\right|^{\gamma/2} \mid \mathbb{Z}^{**}_{n,k}(t_2) - \mathbb{Z}^{**}_{n,k}(t)\right|^{\gamma/2}\right) = 0 \leq (t_2 - t_1)^2$.

Let $t_2 - t_1 \geq 1/n$, then there are 3 possible cases:

(1) $t_2 - t \geq \frac{1}{2n}, t - t_1 \geq \frac{1}{2n}$, then from Cauchy-Bunyakovsky Inequality, $\mathbb{Q} \leq \widetilde{C}(\theta)(t_2 - t)^2(t - t_1)^2 \leq \widetilde{C}(\theta)(t_2 - t_1)^2$.

(2) $t_2 - t \geq \frac{1}{2n}, t - t_1 < \frac{1}{2n}$, then from Cauchy-Bunyakovsky Inequality,

$$\mathbb{Q} \leq \sqrt{\widetilde{C}(\theta)(t_2 - t)^4 \mathbf{E}\left(\frac{\Pi(1) + 1}{\sqrt{\alpha(n)}}\right)^\gamma} \leq \widehat{C}(\theta)(t_2 - t_1)^2.$$

(3) $t_2 - t < \frac{1}{2n}, t - t_1 \geq \frac{1}{2n}$, symmetric to case 2.

So we have (see Billingsley [10], Theorem 13.5) tihgtness of k-th component and therefore tihgtness of all the vector.

Step 4 (approximation of the original process) From the relative compactness of distributions of processes $\left\{\mathbb{Z}^*_{n,k}\right\}_{n\geq n_0, k\geq 1}$ we get that for every pair $\varepsilon > 0, \eta > 0$ there exist $\delta \in (0, 1)$ and $N_1 = N_1(\varepsilon, \eta)$ such that for all $n \geq N_1$

$$\mathbf{P}\left(\sup_{|t-\tau|\leq\delta}\left|\mathbb{Z}^*_{n,k}(\tau) - \mathbb{Z}^*_{n,k}(t)\right| \geq \eta\right) \leq \varepsilon.$$

We have $\mathbf{P}\left(\mathbb{Y}^*_{n,k}(t) = \mathbb{Z}^*_{n,k}(\tau) \mid \Pi(n\tau) = [nt]\right) = 1$. Let N be from Lemma 1 (iii). We remember the designation

$$A(n) = \{\forall t \in [0, 1] \ \ \exists\tau : |\tau - t| \leq \delta, \ \Pi(n\tau) = [nt]\}.$$

Thus for all $n \geq \max(N, N_1)$

$$\mathbf{P}\left(\sup_{0\leq t\leq 1}\left|\mathbb{Y}^*_{n,k}(t) - \mathbb{Z}^*_{n,k}(t)\right| \geq \eta\right) \leq \mathbf{P}\left(\sup_{0\leq t\leq 1}\left|\mathbb{Y}^*_{n,k}(t) - \mathbb{Z}^*_{n,k}(t)\right| \geq \eta, A(n)\right) + \varepsilon$$

$$\leq \mathbf{P}\left(\sup_{|t-\tau|\leq\delta}\left|\mathbb{Z}^*_{n,k}(\tau) - \mathbb{Z}^*_{n,k}(t)\right| \geq \eta\right) + \varepsilon \leq 2\varepsilon.$$

The proof is complete.

For *Childe Harold's Pilgrimage* we have a good match with $\theta = 0.54$, $q_1 = 1$. For *Evgene Onegin* we need some nonzero q_2, we have $\theta = 0.72$, $q_1 = 0.96$, $q_2 = 0.04$.

7. Online Simon modification of the Karlin model

The disadvantage of the Simon model is the linear growth of R_n. We need a power growth with an exponent lesser than 1. Our idea is to use the classical infinite urn model with re-distribution: any ball takes an urn independently with some discrete power law; if the ball is falling in an empty urn then it is re-distributed uniformly on all previous urns with one ball with probability $1 - p$, and it stays in the new urn with probability p.

The model starts from the infinite sequence of empty urns. The first ball takes one of the urns with the integer-valued power law with exponent $1/\beta$, $0 < \beta < 1$. Each next ball takes one of the urns with the same law, independently of previous balls. After this, any next ball, if it is in a new urn, is re-tossed independently, that is, with probability $1 - p$ selects one of the previous urns with one ball at random and joins it, like in the Simon model. All other balls stay in selected urns. This is the online variant of the model from Section 5.

We have a good match for *Childe Harold's Pilgrimage* with $\beta = 0.54$, $p = 0$. We need some nonzero p to correspond the model to *Evgene Onegin*, $\beta = 0.73$, $p = 0.08$.

The simulation shows that statistics $\widehat{\theta}$ and θ^* converge to different limits under these models, in contrast to the elementary urn model. However, the analytical dependencies of these limits on the parameters of the models remain unclear.

8. Discussion

So, the infinite urn model (Karlin model) and the Simon model are the most known models of texts, but they have disadvantages that do not give the ability to simulate the number of unique words correctly. The infinite urn model states too strict conditions for the relative number of unique and different words. The Simon model states a linear growth of the numbers of different and unique words. Its modification by Baur and Bertoin [9] preserves the linear dependence.

We propose three modifications of the Karlin and Simon models. The first one is the offline variant, the Simon model starts after the completion of the infinite urn scheme. We have analytical theorems (SLLN and FCLT) for the process in embedded times of increation of the initial urn process. This model holds any relative number of unique and different words, but we can test the model under the condition of the given initial urn process only.

The second modification involves the compound Poisson process in the infinite urn model. As a variant, any word can be doubled with some positive probability. We prove SLLN and FCLT for the modification. This model can give the decrease (not increase) of the number of different words only. But this decrease is helpful in applications. So we have a simple model that covers the range of parameters that is actual for applications. We have analytical results that can be used for testing the model. On the other side, the model with repeating words is strange.

The third modification is the online variant, the Simon redistribution works at any toss of the Karlin model. Similar to the compound Poisson model, it can give the decrease of the number of unique words only. But, in contrast to the compound Poisson model, we have no analytics (limit theorems) for this modification. The online modification seems to be the most logically supported, but we can study it by simulation only due to the absence of theoretically supported tests.

Acknowledgements The reported study was funded by RFBR and CNRS according to the research project No. 19-51-15001. The authors like to thank Bastien Mallein for fruitful and constructive comments and suggestions, especially for the interpretation of Simon's model in terms of Janson's generalized Polya urn scheme. The authors thank Sergey Foss and Sanjay Ramassamy for helpful discussions. Finally, the authors thank an anonymous referee for careful reading of the paper and a number of very good suggestions that helped to improve the presentation of the results.

References

[1] R.J. Adler, *An introduction to continuity, extrema, and related topics for general Gaussian processes*, Institute of Math. Stat., Hayward, 1990. Zbl 0747.60039

[2] D.J. Aldous, *Probability distributions on cladograms*, In: Aldous, David (ed.) et al., *Random discrete structures. Based on a workshop held November 15-19, 1993 at IMA, University of Minnesota, Minneapolis*, Springer, Berlin, 1996, 1–18. Zbl 0841.92015

[3] D. Aldous, J. Pitman, *The standard additive coalescent*, Ann. Probab., **26**:4 (1998), 1703–1726. Zbl 0936.60064

[4] R.R. Bahadur, *On the number of distinct values in a large sample from an infinite discrete distribution*, Proc. Natl. Inst. Sci. India, Part A, Suppl. II, **26** (1960), 67–75. Zbl 0151.23803

[5] A.D. Barbour, *Univariate approximations in the infinite occupancy scheme*, Alea, **6** (2009), 415–433. MR2576025

[6] A.D. Barbour, A.V. Gnedin, *Small counts in the infinite occupancy scheme*, Electron. J. Probab., **14** (2009), 365–384. Zbl 1189.60048

[7] E. Baur, J. Bertoin, *Cutting edges at random in large recursive trees*, In Crisan, Dan (ed.) et al., *Stochastic analysis and applications 2014. In honour of Terry Lyons. Selected articles based on the presentations at the conference, Oxford, UK, September 23–27, 2013*, Springer Proceedings in Mathematics & Statistics, **100**, Springer, Cham, 2014, 51–76. Zbl 1390.60043

[8] E. Baur, J. Bertoin, 2015. *The fragmentation process of an infinite recursive tree and Ornstein-Uhlenbeck type processes*, Electron. J. Probab., **20** (2015), Paper No. **98**. Zbl 1333.60186

[9] E. Baur, J. Bertoin, *On a two-parameter Yule-Simon distribution*, arXiv:2001.01486, 2020.

[10] P. Billingsley, *Convergence of probability measures.*, Wiley, Chichester, 1999. Zbl 0944.60003

[11] 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. Zbl 1340.62018

[12] M. Chebunin, A. Kovalevskii, *Functional central limit theorems for certain statistics in an infinite urn scheme*, Stat. Probab. Lett., **119** (2016), 344–348. Zbl 1398.60051

[13] M. Chebunin, A. Kovalevskii, *Asymptotically normal estimators for Zipf's law*, Sankhyā, Ser. A, **81**:2 (2019), 482–492. Zbl 1437.62097

[14] M. Chebunin, A. Kovalevskii, *A statistical test for the Zipf's law by deviations from the Heaps' law*, Sib. Èlektron. Mat. Izv., **16** (2019), 1822–1832. Zbl 1433.62060

[15] O. Durieu, Y. Wang, *From infinite urn schemes to decompositions of self-similar Gaussian processes*, Electron. J. Probab., **21** (2015), Paper No. **43**, Zbl 1346.60039

[16] M. Dutko, *Central limit theorems for infinite urn models*, Ann. Probab., **17**:3 (1989), 1255–1263. Zbl 0685.60023

[17] A. Gnedin, B. Hansen, J. Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws*, Probab. Surv., **4** (2007), 146–171. Zbl 1189.60050

[18] F.A. Haight, R.B. Jones, *A probabilistic treatment of qualitative data with special reference, to word association tests*, J. Math. Psychol., **11** (1974), 237–244. Zbl 0287.92011

[19] H.S. Heaps, *Information retrieval. Computational and theoretical aspects*, Academic Press, New York etc., 1978. Zbl 0471.68075

[20] G. Herdan, *Type-token mathematics*, Mouton and Co., 'S-Gravenhage, 1960. Zbl 0163.40904

[21] H.-K. Hwang, S. Janson, *Local limit theorems for finite and infinite urn models*, Ann. Probab., **36**:3 (2008), 992–1022. Zbl 1138.60027

[22] S. Janson, *Functional limit theorems for multitype branching processes and generalized Pólya urns*, Stochastic Processes Appl., **110**:2 (2004), 171–245. Zbl 1075.60109

[23] S. Karlin, *Central limit theorems for certain infinite urn schemes.* J. Math. Mech., **17**:4 (1967), 373–401. Zbl 0154.43701

[24] E.S. Key, *Rare numbers*, J. Theor. Probab., **5**:2 (1992), 375–389. Zbl 0763.60013

[25] E.S. Key, *Divergence rates for the number of rare numbers*, J. Theor. Probab., **9**:2 (1996), 413–428. Zbl 0870.60023

[26] P. Lansky, T. Radill-Weiss, *A generalization of the Yule-Simon model, with special reference to word association tests and neural cell assembly formation*, J. Math. Psychol., **21** (1980), 53–65. Zbl 0443.62078

[27] H.A. Simon, *On a class of skew distribution functions*, Biometrika, **42**:3-4 (1955), 425–440. Zbl 0066.11201

[28] G.U. Yule, *A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S*, Philosophical Transactions of the Royal Society of London. Series B, **213** (1925), 21–87.

[29] N. Zakrevskaya, A. Kovalevskii, *An omega-square statistics for analysis of correspondence of small texts to the Zipf–Mandelbrot law*, in B. Lemeshko (ed) et al., *Applied methods of statistical analysis. Statistical computation and simulation — AMSA'2019*, Proceedings of the International Workshop, NSTU, Novosibirsk, 2019, 488–494.

Mikhail Georgievich Chebunin
Karlsruhe Institute of Technology,
Institute of Stochastics,
Karlsruhe, 76131, Germany
Novosibirsk State University,
1, Pirogova str.,
Novosibirsk, 630090, Russia
Email address: chebuninmikhail@gmail.com

Artyom Pavlovich Kovalevskii
Novosibirsk State Technical University,
20, K. Marksa ave.,
630073, Novosibirsk, Russia
Novosibirsk State University,
1, Pirogova str.,
Novosibirsk, 630090, Russia
Email address: artyom.kovalevskii@gmail.com