ISSN 1342-2804

Research Reports on Mathematical and Computing Sciences

Limiting size index distributions for Zipf-type word frequencies

Satoshi Chida and Naoto Miyoshi

December 2007, B–446

Department of Mathematical and Computing Sciences
Tokyo Institute of Technology

SERIES B: Operations Research

Limiting size index distributions for Zipf-type word frequencies

Satoshi Chida and Naoto Miyoshi*

Department of Mathematical and Computing Sciences
Tokyo Institute of Technology

Abstract

We consider a random linguistic model where words are placed randomly and sequentially in a given text or corpus. The word frequency follows the Zipf-type distribution; that is, the probability with which the ith most popular word occurs is asymptotically proportional to $1/i^{\alpha}$, $\alpha > 0$. In this model, we derive the limiting distributions of size indices, where the size index of degree k at time t represents the number of distinct words appearing exactly k times during $(0, t]$. While the past studies only treated the case where the parameter α of the Zipf-type distribution is greater than unity, we here consider the case of $\alpha \leq 1$ as well as $\alpha > 1$. We first investigate the limiting size index distributions for the independent word occurrence model and then extend the derived results to the case where the word occurrences are generally dependent. Simulation experiments demonstrate not only that our analysis is valid but also that the derived limiting distributions well approximate the size index distributions for relatively short texts.

Keywords: Random linguistic model, size indices, limiting distribution, Zipf-type distribution.

*Corresponding author: 2-12-1-W8-52 Ookayama, Tokyo 152-8552, Japan. E-mail: miyoshi@is.titech.ac.jp

1 Introduction

We consider a random linguistic model where words are placed randomly and sequentially in a given text or corpus. One of important criteria characterizing the feature of texts and corpora is the distribution of *size indices*, which are also called *frequency spectra* or *frequencies of frequencies* (see, e.g., Baayen [1] and Sibuya [10]). Now, let m ($\in \overline{\mathbb{N}} = \mathbb{N} \cup \{+\infty\}$) denote the size of vocabulary; that is, the number of elements in a given set of words, and let $N_i^{(m)}(t)$, $i = 1, \ldots, m$, $t \geq 0$, denote the number of word i appearing in the text during an observation period $(0, t]$. Then, the size index $S_k^{(m)}(t)$ of degree k ($= 1, 2, \ldots$) at time t is defined as $S_k^{(m)}(t) = \sum_{i=1}^{m} 1_{\{N_i^{(m)}(t)=k\}}$, where 1_E denotes the indicator function for event E; that is, $S_k^{(m)}(t)$ denotes the number of distinct words appearing exactly k times in the text during $(0, t]$. The size indices are indeed important for applications to, for example, efficient coding or data compression of computerized texts. In this paper, we investigate the limiting distributions of size indices with respect to an increase of the observation period (and an increase of the vocabulary size when $m < \infty$); that is, we derive the almost sure limits of $S_k^{(\infty)}(t)/\overline{S}_1^{(\infty)}(t)$, $k = 1, 2, \ldots$, as $t \to \infty$ when $m = +\infty$ and those of $S_k^{(m)}(t(m))/m$, $k = 0, 1, 2, \ldots$, as $m \to \infty$ when $m < \infty$, where $\overline{S}_k^{(m)}(t) = \sum_{l \geq k} S_l^{(m)}(t)$, $S_0^{(m)}(t) = m - \overline{S}_1^{(m)}(t)$ and $t(m)$ is some function of m satisfying $t(m) \to \infty$ as $m \to \infty$.

The study concerning the asymptotics of the size indices dates back to 1960–70's. Karlin [3] considered the case where the vocabulary size is infinite and the occurrences of words are i.i.d. for both the discrete-time model; that is, a word is placed at every time unit, and the Poisson embedded model; that is, words are placed according to a homogeneous Poisson process. Among many other results, he derived the asymptotics of $\mathrm{E}S_k^{(\infty)}(t)$, $k = 1, 2, \ldots$, as $t \to \infty$ and also showed that $S_k^{(\infty)}(t)/\mathrm{E}S_k^{(\infty)}(t) \to 1$ almost surely as $t \to \infty$, where E denotes the expectation. While it is easy to derive the limit of the empirical size index distribution $S_k^{(\infty)}(t)/\overline{S}_1^{(\infty)}(t)$, $k = 1, 2, \ldots$, as $t \to \infty$ from Karlin's results, Rouault [7] further extended it to the case where the words occurrences are Markov dependent under the assumption that the word frequency distribution follows the generalized Zipf-type law; that is, the distribution with regularly varying tail.

From the past, extensive studies have developed lexical models with Zipf-like law for word frequency (see, e.g., [1] and references therein). We then employ the Zipf-type distribution as the model of word frequency; that is, the probability with which word i is chosen is asymptotically proportional to $1/i^{\alpha}$ with $\alpha > 0$ for large i and m. Note here that, if the vocabulary size m is infinite, then the parameter α of the Zipf-type distribution must be greater than unity. In this paper, we extend the results in [3, 7] mentioned above towards two directions. In one direction, we consider the case of $\alpha \leq 1$ and $m < \infty$, where the vocabulary size m increases together with the observation period. It is illustrated that, by determining an appropriate function $t(m)$, $m \in \mathbb{N}$, we can obtain the limit of the empirical size index distribution $S_k^{(m)}(t(m))/m$, $k = 0, 1, 2, \ldots$, as $m \to \infty$. As a byproduct, we also find that such a size index distribution offers a natural solution to the so-called zero-frequency problem (see, e.g., Witten and Bell [11]) in the particular case of the Zipf-type word frequency with $\alpha \leq 1$. The other extension is that the word occurrences are generally dependent while their marginal distributions are a common Zipf-type one, where we show that the results derived for the independent word occurrence model are still valid for the dependent word occurrence model under some additional conditions. The model we here consider is the Poisson embedded one; that is, words are placed according to a homogeneous Poisson process with intensity 1, and at each point of the Poisson process, a word is chosen according to the given Zipf-type probability.

The organization of the paper is as follows. In the next section, we consider the independent word occurrence model, where we first describe the model and review the existing results for the case of $\alpha > 1$ and $m = +\infty$. We then consider the cases of $\alpha < 1$ and $\alpha = 1$ with $m < +\infty$ and derive the limiting

size index distributions in the form of $\lim_{m\to\infty} S_k^{(m)}(t(m))/m$, $k = 0, 1, 2, \ldots$, for some $t(m)$, $m \in \mathbb{N}$. In Section 3, we extend the model as the word occurrences are dependent in some general sense and verify that all results in Section 2 are still valid under some additional conditions. The derived results are validated through simulation experiments in Section 4, where we also find that the limiting distributions well approximate the size index distributions for relatively short texts.

2 Independent word occurrence model

Throughout this and the next sections, we suppose that all random elements are defined on a common probability space $(\Omega, \mathcal{F}, \mathrm{P})$. In the analysis, we use the following standard notation; that is, for any two real functions $f(x)$ and $g(x)$ for $x \in \mathbb{R}$, $f(x) \sim g(x)$ as $x \to a$ stands for $\lim_{x\to a} f(x)/g(x) = 1$, where a is possibly $+\infty$.

2.1 Model description

Let $\{N(t)\}_{t\geq 0}$ denote a homogeneous Poisson process with intensity 1, where $N(t)$, $t \geq 0$, represents the number of points of the Poisson process during $(0, t]$. At each point of $\{N(t)\}_{t\geq 0}$, a word is chosen randomly from the vocabulary set $\{1, \ldots, m\}$ when $m < +\infty$, or $\{1, 2, \ldots\}$ when $m = +\infty$, and is placed in the text. Let W_n, $n \in \mathbb{N}$, denote the random variable representing the word which is chosen at the nth point of $\{N(t)\}_{t\geq 0}$. We suppose that W_n, $n \in \mathbb{N}$, are mutually independent and are also independent of the Poisson process $\{N(t)\}_{t\geq 0}$. The probability with which word i $(= 1, \ldots, m)$ is chosen is denoted by $p_i^{(m)} = \mathrm{P}(W_1 = i)$, where $p_i^{(m)} \geq 0$, $i = 1, \ldots, m$, and $p_1^{(m)} + \cdots + p_m^{(m)} = 1$. Let $N_i^{(m)}(t)$, $i = 1, \ldots, m$, denote the number of word i appearing in $(0, t]$, so that $\sum_{i=1}^m N_i^{(m)}(t) = N(t)$, $t \geq 0$. By the fundamental property of Poisson processes, $N_i^{(m)}(t)$, $i = 1, \ldots, m$, are mutually independent and also follow the Poisson distributions with mean $p_i^{(m)}\, t$; that is,

$$\mathrm{P}\big(N_1^{(m)}(t) = k_1, \ldots, N_m^{(m)}(t) = k_m\big) = \prod_{i=1}^m \frac{(p_i^{(m)}\, t)^{k_i}}{k_i!}\, e^{-p_i^{(m)} t}, \quad t \geq 0,\, k_1, \ldots, k_m \in \mathbb{Z}_+.$$

We assume that the word frequency distribution $\boldsymbol{p}^{(m)} = (p_1^{(m)}, \ldots, p_m^{(m)})$ is Zipf-type; that is, $p_i^{(m)}$ is asymptotically proportional to $1/i^\alpha$, $\alpha > 0$, for large i and m. Within this setting, we investigate below the limiting distributions of size indices $S_k^{(m)}(t) = \sum_{i=1}^m 1_{\{N_i^{(m)}(t)=k\}}$, $k = 1, 2, \ldots$, for the cases of $\alpha > 1$, $\alpha < 1$ and $\alpha = 1$ separately.

2.2 Case of $\alpha > 1$

In this subsection, we review the results for the independent word occurrence model with Zipf-type frequency in the case where the vocabulary size m is infinite and the parameter α of the Zipf-type distribution is greater than unity. The results are obtained directly from [3, 7]. We here suppress the superscript "(∞)" and write, for example, p_i, $N_i(t)$ for $p_i^{(\infty)}$, $N_i^{(\infty)}(t)$ and so on. In this case, the Zipf-type word frequency distribution $\boldsymbol{p} = (p_1, p_2, \ldots)$ is provided as the following.

Assumption 1 $p_i \sim c/i^\alpha$ as $i \to \infty$ with $\alpha > 1$ and $c > 0$. Namely, for any $\epsilon > 0$, there exists an integer $i_\epsilon > 0$ such that, for all $i \geq i_\epsilon$, inequality $(1 - \epsilon)\, c/i^\alpha \leq p_i \leq (1 + \epsilon)\, c/i^\alpha$ holds.

As defined above, let $S_k(t)$, $t \geq 0$, denote the size index of degree k $(= 1, 2, \ldots)$ at time t (≥ 0); that is, the number of distinct words appearing exactly k times during $(0, t]$. Let also $\overline{S}_k(t) = \sum_{l\geq k} S_l(t)$; that is, the number of distinct words appearing at least k times during $(0, t]$. Then, Karlin [3] derived the following result in more general form.

Proposition 2.1 *Under Assumption 1, we have for* $k = 1, 2, \ldots$,

$$\mathrm{E}\overline{S}_k(t) \sim c^{1/\alpha}\, t^{1/\alpha} \Big[\Gamma\Big(1 - \frac{1}{\alpha}\Big) - \sum_{l=1}^{k-1} \frac{1}{\alpha\, l!}\, \Gamma\Big(l - \frac{1}{\alpha}\Big)\Big] \quad \textit{as } t \to \infty, \tag{2.1}$$

where $\sum_{l=1}^{0} \cdot = 0$ *conventionally and* Γ *denotes Euler's Gamma function; that is,* $\Gamma(x) = \int_0^\infty e^{-u}\, u^{x-1}\, \mathrm{d}u$.

Applying $S_k(t) = \overline{S}_k(t) - \overline{S}_{k+1}(t)$ a.s. in Proposition 2.1, we can readily obtain the limits of $\mathrm{E}S_k(t)/\mathrm{E}\overline{S}_1(t)$, $k = 1, 2, \ldots$, as $t \to \infty$ under Assumption 1. Furthermore, Karlin [3] proved that $S_k(t)/\mathrm{E}S_k(t) \to 1$ a.s. as $t \to \infty$. These results immediately yield the following.

Proposition 2.2 *Under Assumption 1, we have for* $k = 1, 2, \ldots$,

$$\lim_{t\to\infty} \frac{S_k(t)}{\overline{S}_1(t)} = \frac{\Gamma(k - 1/\alpha)}{\alpha\, k!\, \Gamma(1 - 1/\alpha)} = \frac{1}{\alpha\, k} \prod_{l=1}^{k-1} \Big(1 - \frac{1}{\alpha\, l}\Big) \quad \textit{a.s.} \tag{2.2}$$

Let $\Psi_k(\alpha)$, $\alpha > 1$, $k = 1, 2, \ldots$, denote the right-hand side of (2.2). We can check that $\Psi_k(\alpha)$, $k = 1, 2, \ldots$, gives a proper distribution on $\mathbb{N}$ with $\Psi_k(\alpha) \sim k^{-1-1/\alpha}/[\alpha\, \Gamma(1-1/\alpha)]$ as $k \to \infty$. Rouault [7] provided the same formula for the discrete-time model and further extended as it is also valid when the word occurrences are Markov dependent. The same distribution was derived by Sibuya [9] independently in another problem, so that we refer to the distribution given by $\Psi_k(\alpha)$, $k = 1, 2, \ldots$, as the Karlin-Rouault-Sibuya (KRS) distribution. Propositions 2.1 and 2.2 are extended to the case where the word occurrences are generally dependent in Section 3.

2.3 Case of $\alpha < 1$

We here consider the case where the parameter α of the Zipf-type distribution is less than unity and the vocabulary size m is finite. We derive the limiting size index distribution with respect to an increase of m together with the observation period. The word frequency distribution is supposed to satisfy the following.

Assumption 2 For any $\epsilon > 0$, there exists an integer $i_\epsilon > 0$ such that, for all m and i satisfying $m \geq i \geq i_\epsilon$, inequality $(1 - \epsilon)\, c/(m^{1-\alpha}\, i^\alpha) \leq p_i^{(m)} \leq (1 + \epsilon)\, c/(m^{1-\alpha}\, i^\alpha)$ holds with $\alpha \in (0, 1)$ and $c > 0$.

Note that Assumption 2 represents the Zipf-type distribution with $\alpha < 1$. Indeed, if $p_i^{(m)} = c_m/i^\alpha$ for $i = 1, \ldots, m$ with the normalization constant c_m, we then have $c_m = (\sum_{i=1}^m 1/i^\alpha)^{-1} \sim (1-\alpha)/m^{1-\alpha}$ as $m \to \infty$ and Assumption 2 is fulfilled with $c = 1 - \alpha$. Under this assumption, we first derive the asymptotic result for the expected size indices, which attracts an independent interest and is also exploited even in the extension to the dependent word occurrence model in the next section.

Lemma 2.1 *Under Assumption 2, we have for any fixed constant* $\delta > 0$ *and* $k = 1, 2, \ldots$,

$$\mathrm{E}\overline{S}_k^{(m)}\Big(\frac{\delta\, m}{c}\Big) \sim m \Big[1 - \sum_{l=0}^{k-1} \frac{\delta^{1/\alpha}}{\alpha\, l!}\, \Gamma\Big(l - \frac{1}{\alpha},\, \delta\Big)\Big] \quad \textit{as } m \to \infty, \tag{2.3}$$

where $\Gamma(x, y)$, $y > 0$, *is the incomplete Gamma function; that is,* $\Gamma(x, y) = \int_y^\infty e^{-u}\, u^{x-1}\, \mathrm{d}u$ *(see, e.g., Davis [2]).*

Proof: Since $\mathrm{E}\overline{S}_k^{(m)}(t) = \sum_{i=1}^m \mathrm{P}(N_i^{(m)}(t) \geq k)$ and $N_i^{(m)}(t)$, $i = 1, \ldots, m$, are Poisson random variables with mean $p_i^{(m)}\, t$, we have

$$\mathrm{E}\overline{S}_k^{(m)}(t) = \sum_{i=1}^m \Big[1 - \sum_{l=0}^{k-1} \frac{(p_i^{(m)}\, t)^l}{l!}\, e^{-p_i^{(m)} t}\Big]. \tag{2.4}$$

Here, under Assumption 2, for any $\epsilon \in (0,1)$ and sufficiently large m,

$$\begin{aligned}
\mathrm{E}\overline{S}_k^{(m)}(t) &\le m - \sum_{i=i_\epsilon}^{m} \sum_{l=0}^{k-1} \frac{1}{l!} \Big(\frac{(1-\epsilon)\, c\, t}{m^{1-\alpha}\, i^\alpha} \Big)^l \exp\Big(- \frac{(1+\epsilon)\, c\, t}{m^{1-\alpha}\, i^\alpha} \Big) \\
&\le m - \sum_{l=0}^{k-1} \frac{1}{l!} \int_{i_\epsilon}^{m} \Big(\frac{(1-\epsilon)\, c\, t}{m^{1-\alpha}\, x^\alpha} \Big)^l \exp\Big(- \frac{(1+\epsilon)\, c\, t}{m^{1-\alpha}\, x^\alpha} \Big)\, \mathrm{d}x + \sum_{l=1}^{k-1} \frac{1}{l!} \Big(\frac{(1-\epsilon)\, l}{(1+\epsilon)\, e} \Big)^l, \qquad (2.5)
\end{aligned}$$

where the last term arises from the possibility that the unique maximum of the integrand lies in (i_ϵ, m) for $l = 1, \ldots, k-1$. Similarly, for any $\epsilon \in (0,1)$ and sufficiently large m,

$$\begin{aligned}
\mathrm{E}\overline{S}_k^{(m)}(t) &\ge m - i_\epsilon - \sum_{i=i_\epsilon}^{m} \sum_{l=0}^{k-1} \frac{1}{l!} \Big(\frac{(1+\epsilon)\, c\, t}{m^{1-\alpha}\, i^\alpha} \Big)^l \exp\Big(- \frac{(1-\epsilon)\, c\, t}{m^{1-\alpha}\, i^\alpha} \Big) \\
&\ge m - i_\epsilon - \sum_{l=0}^{k-1} \frac{1}{l!} \int_{i_\epsilon}^{m+1} \Big(\frac{(1+\epsilon)\, c\, t}{m^{1-\alpha}\, x^\alpha} \Big)^l \exp\Big(- \frac{(1-\epsilon)\, c\, t}{m^{1-\alpha}\, x^\alpha} \Big)\, \mathrm{d}x - \sum_{l=1}^{k-1} \frac{1}{l!} \Big(\frac{(1+\epsilon)\, l}{(1-\epsilon)\, e} \Big)^l. \qquad (2.6)
\end{aligned}$$

Now, let $t = \delta\, m/c$ with a constant $\delta > 0$ and change the variable to $u = (1+\epsilon)\, \delta\, m^\alpha / x^\alpha$ in (2.5) and $u = (1-\epsilon)\, \delta\, m^\alpha / x^\alpha$ in (2.6) respectively. We then have

$$\mathrm{E}\overline{S}_k^{(m)}\Big(\frac{\delta\, m}{c} \Big) \le m - \sum_{l=0}^{k-1} \frac{(1-\epsilon)^l\, \delta^{1/\alpha}\, m}{(1+\epsilon)^{l-1/\alpha}\, \alpha\, l!} \int_{(1+\epsilon)\delta}^{(1+\epsilon)\delta m^\alpha / {i_\epsilon}^\alpha} u^{l-1/\alpha-1}\, e^{-u}\, \mathrm{d}u + o(m), \qquad (2.7)$$

$$\mathrm{E}\overline{S}_k^{(m)}\Big(\frac{\delta\, m}{c} \Big) \ge m - \sum_{l=0}^{k-1} \frac{(1+\epsilon)^l\, \delta^{1/\alpha}\, m}{(1-\epsilon)^{l-1/\alpha}\, \alpha\, l!} \int_{(1-\epsilon)\delta m^\alpha / (m+1)^\alpha}^{(1-\epsilon)\delta m^\alpha / {i_\epsilon}^\alpha} u^{l-1/\alpha-1}\, e^{-u}\, \mathrm{d}u + o(m). \qquad (2.8)$$

Thus, dividing both sides of (2.7) and (2.8) by m, taking $m \to \infty$ and finally $\epsilon \downarrow 0$, we obtain (2.3). □

Since $S_0^{(m)} = m - \overline{S}_1^{(m)}(t)$ and $S_k^{(m)}(t) = \overline{S}_k^{(m)}(t) - \overline{S}_{k+1}^{(m)}(t)$ a.s. for $k = 1, 2, \ldots$, we can easily derive the limits of $\mathrm{E}S_k^{(m)}(\delta m/c)/m$, $k = 0, 1, 2, \ldots$, as $m \to \infty$ for any constant $\delta > 0$ from (2.3) in Lemma 2.1; that is, Lemma 2.1 yields the limit of the expected size index distribution for $\alpha < 1$. We now prove its almost sure convergence.

Theorem 2.1 *Under Assumption 2, we have for any fixed constant $\delta > 0$ and $k = 0, 1, 2, \ldots$,*

$$\lim_{m\to\infty} \frac{1}{m}\, S_k^{(m)}\Big(\frac{\delta\, m}{c} \Big) = \frac{\delta^{1/\alpha}}{\alpha\, k!}\, \Gamma\Big(k - \frac{1}{\alpha},\, \delta \Big) \quad \textit{a.s.} \qquad (2.9)$$

Remark 1 Let $\Phi_k(\alpha, \delta)$ denote the right-hand side of (2.9). Then, we can confirm that $\Phi_k(\alpha, \delta)$, $k = 0, 1, 2, \ldots$, gives a proper distribution on $\mathbb{Z}_+$ with $\Phi_k(\alpha, \delta) \sim (\delta^{1/\alpha}/\alpha)\, k^{-1-1/\alpha}$ as $k \to \infty$ even for $\alpha > 0$ (extending the range of α). Furthermore, we can show that, for $\alpha > 0$, $\Phi_k(\alpha, \delta)$ degenerates as

$$\lim_{\delta \downarrow 0} \Phi_k(\alpha, \delta) = \begin{cases} 1, & k = 0, \\ 0, & k = 1, 2, \ldots. \end{cases} \qquad (2.10)$$

These results are verified in Appendix.

Proof: Since $\overline{S}_k^{(m)}(t) = \sum_{i=1}^m 1_{\{N_i^{(m)}(t) \ge k\}}$ a.s. and $N_i^{(m)}(t)$, $i = 1, \ldots, m$, are mutually independent, the Chernoff-Hoeffding bound for the sum of 0-1 independent random variables (see, e.g., Motwani and Raghavan [4, Chapter 4]) implies that, for any $\epsilon > 0$, there exists a $\theta_\epsilon > 0$ such that

$$\mathrm{P}\Big(\big| \overline{S}_k^{(m)}(t) - \mathrm{E}\overline{S}_k^{(m)}(t) \big| > \epsilon\, \mathrm{E}\overline{S}_k^{(m)}(t) \Big) \le 2\, e^{-\theta_\epsilon \mathrm{E}\overline{S}_k^{(m)}(t)}. \qquad (2.11)$$

Furthermore, Lemma 2.1 says that $\mathrm{E}\overline{S}_k^{(m)}(\delta\, m/c) = \Theta(m)$ as $m \to \infty$ under Assumption 2, so that (2.11) leads to

$$\sum_{m=1}^{\infty} \mathrm{P}\biggl(\biggl| \frac{\overline{S}_k^{(m)}(\delta\, m/c)}{\mathrm{E}\overline{S}_k^{(m)}(\delta\, m/c)} - 1 \biggr| > \epsilon \biggr) < \infty. \tag{2.12}$$

Hence, the Borel-Cantelli lemma implies that $\overline{S}_k^{(m)}(\delta\, m/c)/\mathrm{E}\overline{S}_k^{(m)}(\delta\, m/c) \to 1$ a.s. as $m \to \infty$, and applying Lemma 2.1 again, we have for $k = 1, 2, \ldots$,

$$\lim_{m\to\infty} \frac{1}{m}\, \overline{S}_k^{(m)}\Bigl(\frac{\delta\, m}{c}\Bigr) = 1 - \sum_{l=0}^{k-1} \frac{\delta^{1/\alpha}}{\alpha\, l!}\, \Gamma\Bigl(l - \frac{1}{\alpha},\, \delta\Bigr) \quad \text{a.s.},$$

which readily leads to (2.9). □

To compare with the result for the case of $\alpha > 1$, we here present the limiting size index distribution in the form of $\lim_{m\to\infty} S_k^{(m)}(\delta\, m/c)/\overline{S}_1^{(m)}(\delta\, m/c)$, $k = 1, 2, \ldots$.

Corollary 2.1 *Under Assumption 2, we have for any fixed constant $\delta > 0$ and $k = 1, 2, \ldots$,*

$$\lim_{m\to\infty} \frac{S_k^{(m)}(\delta\, m/c)}{\overline{S}_1^{(m)}(\delta\, m/c)} = \frac{\Gamma\bigl(k - 1/\alpha,\, \delta\bigr)}{\alpha\, k!\, \bigl[\Gamma\bigl(1 - 1/\alpha,\, \delta\bigr) + \delta^{-1/\alpha}\,(1 - e^{-\delta})\bigr]} \quad \textit{a.s.} \tag{2.13}$$

Remark 2 Let $\Psi_k(\alpha, \delta)$ denote the right-hand side of (2.13). Then, it is clear from Remark 1 that $\Psi_k(\alpha, \delta)$, $k = 1, 2, \ldots$, also gives a proper distribution on $\mathbb{N}$ with $\Psi_k(\alpha, \delta) = \Theta(k^{-1-1/\alpha})$ as $k \to \infty$ even for $\alpha > 0$ (extending the range of α). Furthermore, we can show that, when $\alpha > 1$, $\lim_{\delta\downarrow 0} \Psi_k(\alpha, \delta)$, $k = 1, 2, \ldots$, reduces to the KRS distribution given on the right-hand side of (2.2), but when $\alpha \in (0, 1]$, it degenerates as

$$\lim_{\delta\downarrow 0} \Psi_k(\alpha, \delta) = \begin{cases} 1, & k = 1, \\ 0, & k = 2, 3, \ldots. \end{cases} \tag{2.14}$$

This is verified in Appendix.

2.4 Case of $\alpha = 1$

In the final subsection for the independent word occurrence model, we consider the case of $\alpha = 1$ and $m < \infty$. The word frequency distribution is given by the following.

Assumption 3 For any $\epsilon > 0$, there exists an integer $i_\epsilon > 0$ such that, for all m and i satisfying $m \geq i \geq i_\epsilon$, inequality $(1 - \epsilon)\, c/(i \ln m) \leq p_i^{(m)} \leq (1 + \epsilon)\, c/(i \ln m)$ holds with $c > 0$.

Similar to Assumption 2 in the preceding subsection, Assumption 3 represents the Zipf-type distribution with $\alpha = 1$. Indeed, if $p_i^{(m)} = c_m/i$ for $i = 1, \ldots, m$ with the normalization constant c_m, we have $c_m = (\sum_{i=1}^m 1/i)^{-1} \sim 1/\ln m$ as $m \to \infty$ and Assumption 3 is fulfilled with $c = 1$. As in the preceding subsection, we first provide the asymptotics of the expected size indices.

Lemma 2.2 *Under Assumption 3, for any fixed constant $\delta > 0$ and $k = 1, 2, \ldots$,*

$$\mathrm{E}\overline{S}_k^{(m)}\Bigl(\frac{\delta\, m\, \ln m}{c}\Bigr) \sim m\, \Bigl[1 - \sum_{l=0}^{k-1} \frac{\delta}{l!}\, \Gamma(l - 1,\, \delta)\Bigr] \quad \textit{as } m \to \infty. \tag{2.15}$$

Note here that the right-hand side of (2.15) in Lemma 2.2 is just the version of $\alpha = 1$ on the right-hand side of (2.3) in Lemma 2.1 while the functions $t = t(m)$, $m \in \mathbb{N}$, are different for the two cases.

Proof: The proof is similar to that of Lemma 2.1. The differences are in that, under Assumption 3, inequalities (2.5) and (2.6) are respectively replaced by

$$\mathrm{E}\overline{S}_k^{(m)}(t) \le m - \sum_{l=0}^{k-1} \frac{1}{l!} \int_{i_\epsilon}^{m} \Big(\frac{(1-\epsilon)\, c\, t}{x \, \ln m} \Big)^l \exp\Big(- \frac{(1+\epsilon)\, c\, t}{x \, \ln m} \Big) \, \mathrm{d}x + \sum_{l=1}^{k-1} \frac{1}{l!} \Big(\frac{(1-\epsilon)\, l}{(1+\epsilon)\, e} \Big)^l, \tag{2.16}$$

$$\mathrm{E}\overline{S}_k^{(m)}(t) \ge m - i_\epsilon - \sum_{l=0}^{k-1} \frac{1}{l!} \int_{i_\epsilon}^{m+1} \Big(\frac{(1+\epsilon)\, c\, t}{x \, \ln m} \Big)^l \exp\Big(- \frac{(1-\epsilon)\, c\, t}{x \, \ln m} \Big) \, \mathrm{d}x - \sum_{l=1}^{k-1} \frac{1}{l!} \Big(\frac{(1+\epsilon)\, l}{(1-\epsilon)\, e} \Big)^l. \tag{2.17}$$

Thus, letting $t = (\delta\, m \, \ln m)/c$ with $\delta > 0$ and changing the variable to $u = (1+\epsilon)\, \delta\, m/x$ in (2.16) and $u = (1-\epsilon)\, \delta\, m/x$ in (2.17) respectively, we have

$$\mathrm{E}\overline{S}_k^{(m)}\Big(\frac{\delta\, m \, \ln m}{c} \Big) \le m - \sum_{l=0}^{k-1} \frac{(1-\epsilon)^l\, \delta\, m}{(1+\epsilon)^{l-1}\, l!} \int_{(1+\epsilon)\delta}^{(1+\epsilon)\delta m/i_\epsilon} u^{l-2}\, e^{-u} \, \mathrm{d}u + o(m),$$

$$\mathrm{E}\overline{S}_k^{(m)}\Big(\frac{\delta\, m \, \ln m}{c} \Big) \ge m - \sum_{l=0}^{k-1} \frac{(1+\epsilon)^l\, \delta\, m}{(1-\epsilon)^{l-1}\, l!} \int_{(1-\epsilon)\delta m/(m+1)}^{(1-\epsilon)\delta m/i_\epsilon} u^{l-2}\, e^{-u} \, \mathrm{d}u + o(m).$$

Hence, dividing the both sides by m, taking $m \to \infty$ and then $\epsilon \downarrow 0$, we obtain (2.15). □

Applying Lemma 2.2, a similar procedure to the proof of Theorem 2.1 yields the following.

Theorem 2.2 *Under Assumption 3, for any fixed constant $\delta > 0$ and $k = 0, 1, 2, \ldots$,*

$$\lim_{m\to\infty} \frac{1}{m}\, S_k^{(m)}\Big(\frac{\delta\, m \, \ln m}{c} \Big) = \frac{\delta}{k!}\, \Gamma(k-1,\, \delta) \quad a.s. \tag{2.18}$$

Proof: The difference from the proof of Theorem 2.1 is just that, applying Lemma 2.2 under Assumption 3 (instead of Lemma 2.1 under Assumption 2), (2.12) is replaced by

$$\sum_{m=1}^{\infty} \mathrm{P}\Bigg(\Big| \frac{\overline{S}_k^{(m)}(\delta\, m \, \ln m/c)}{\mathrm{E}\overline{S}_k^{(m)}(\delta\, m \, \ln m/c)} - 1 \Big| > \epsilon \Bigg) < \infty.$$

The remaining procedure is quite the same. □

Remark 3 As seen in Remark 1, the right-hand side of (2.18) in Theorem 2.2 gives a proper distribution on $\mathbb{Z}_+$. What is interesting in (2.9) and (2.18) is that they assign the positive probability to the words which never emerge in the text. This suggests that the distributions given on the right-hand sides of (2.9) and (2.18) offer a natural solution to the so-called zero-frequency problem (see, e.g., [11]) in the particular case of the Zipf-type word frequency with $\alpha \le 1$.

Even in the case of $\alpha = 1$, of course, the following holds.

Corollary 2.2 *Under Assumption 3, we have for any fixed constant $\delta > 0$ and $k = 1, 2, \ldots$,*

$$\lim_{m\to\infty} \frac{S_k^{(m)}\big((\delta\, m \, \ln m)/c\big)}{\overline{S}_1^{(m)}\big((\delta\, m \, \ln m)/c\big)} = \frac{\Gamma(k-1,\, \delta)}{k!\, \big[\Gamma(0,\, \delta) + \delta^{-1}\, (1 - e^{-\delta})\big]} \quad a.s. \tag{2.19}$$

3 Dependent word occurrence model

We here extend the results in the preceding section to the dependent word occurrence model. We will find that the derived limiting size index distributions are still valid even in the case where the word occurrences are dependent in some general sense.

3.1 Model description

As in the independent word occurrence model, words are placed according to a homogeneous Poisson process with intensity 1, but the choice of a word at each point of the Poisson process is governed by the following doubly stochastic structure. Let $\{\boldsymbol{q}^{(m)}(t)\}_{t\ge 0}$ denote a stochastic process on $[0,1]^m$ independent of the Poisson process $\{N(t)\}_{t\ge 0}$ such that $\boldsymbol{q}^{(m)}(t) = (q_1^{(m)}(t), \ldots, q_m^{(m)}(t))$ satisfies $\sum_{i=1}^m q_i^{(m)} = 1$ a.s. for $t \ge 0$. We suppose that $\mathrm{E}q_i^{(m)}(t) = p_i^{(m)}$ for $i = 1, \ldots, m$ and $t \ge 0$. Let T_n, $n = 1, 2, \ldots$, denote the nth point of the Poisson process $\{N(t)\}_{t\ge 0}$. Then, once $\boldsymbol{q}^{(m)}(T_n)$ is given at time T_n, $n = 1, 2, \ldots$, the word W_n is conditionally independent of $\{W_l\}_{l\ne n}$ and $\{\boldsymbol{q}^{(m)}(t)\}_{t\ne T_n}$, and word i $(= 1, \ldots, m)$ is chosen with conditional probability $q_i^{(m)}(T_n) = \mathrm{P}(W_n = i \mid \boldsymbol{q}^{(m)}(T_n))$. Within this setting, each $\{N_i^{(m)}(t)\}_{t\ge 0}$, $i = 1, \ldots, m$, is a Cox (doubly stochastic Poisson) process with random intensity process $\{q_i^{(m)}(t)\}_{t\ge 0}$, and $N_i^{(m)}(t)$, $i = 1, \ldots, m$, are mutually conditionally independent given $\{\boldsymbol{q}^{(m)}(s)\}_{s\le t}$; that is,

$$\mathrm{P}\big(N_1^{(m)}(t) = k_1, \ldots, N_m^{(m)}(t) = k_m \mid \boldsymbol{q}^{(m)}(s), s \le t\big) = \prod_{i=1}^m \frac{\big[Q_i^{(m)}(t)\big]^{k_i}}{k_i!}\, e^{-Q_i^{(m)}(t)} \quad \text{a.s.}, \tag{3.1}$$

where $Q_i^{(m)}(t) = \int_0^t q_i^{(m)}(s)\,\mathrm{d}s$ for $i = 1, \ldots, m$.

3.2 Case of $\alpha > 1$

We here extend Propositions 2.1 and 2.2 in Section 2.2 to the dependent word occurrence model. As in Section 2.2, we suppress the superscript "(∞)" and write, for example, $\boldsymbol{q}(t)$ for $\boldsymbol{q}^{(\infty)}(t)$ and so on. Assuming that $\{\boldsymbol{q}(t)\}_{t\ge 0}$ is ergodic, the ergodic theorem implies that $Q_i(t)/t \to p_i$ a.s. as $t \to \infty$ for $i = 1, 2, \ldots$. Now, for any $t > 0$ and $\epsilon > 0$, we define $A_\epsilon(t) \in \mathcal{F}$ such that

$$A_\epsilon(t) = \left\{\omega \in \Omega \;\middle|\; \sup_{i\in\mathbb{N}} \left|\frac{Q_i(\omega, t)}{p_i\, t} - 1\right| \le \epsilon\right\}. \tag{3.2}$$

Note that the ergodicity implies $\mathrm{P}(A_\epsilon(t)) \to 1$ as $t \to \infty$. We first show the asymptotics of the expected size indices, where some convergence speed of $\mathrm{P}\big(A_\epsilon(t)^c\big) \to 0$ as $t \to \infty$ is required.

Lemma 3.1 *Under Assumption 1, if* $\mathrm{P}(A_\epsilon(t)^c) = o(t^{-1+1/\alpha})$*, then (2.1) holds for* $k = 1, 2, \ldots$.

Proof: We first consider the case of $k = 1$; that is, we verify that, under the condition of the lemma,

$$\mathrm{E}\overline{S}_1(t) \sim c^{1/\alpha}\, t^{1/\alpha}\, \Gamma\Big(1 - \frac{1}{\alpha}\Big) \quad \text{as } t \to \infty. \tag{3.3}$$

Since $\mathrm{E}(\overline{S}_1(t) \mid \boldsymbol{q}(s), s \le t) = \sum_{i=1}^\infty \mathrm{P}(N_i(t) \ge 1 \mid q_i(s), s \le t)$ a.s., we have from (3.1) that

$$\mathrm{E}(\overline{S}_1(t) \mid \boldsymbol{q}(s), s \le t) = \sum_{i=1}^\infty \big[1 - e^{Q_i(t)}\big] \quad \text{a.s.} \tag{3.4}$$

Here, for any $\epsilon \in (0,1)$ and $i = 1, 2, \ldots$, we have $(1-\epsilon)\, p_i\, t \le Q_i(t) \le (1+\epsilon)\, p_i\, t$ on $A_\epsilon(t)$ in (3.2), and thus, under Assumption 1,

$$\begin{aligned}
\mathrm{E}(\overline{S}_1(t) \mid \boldsymbol{q}(s), s \le t) &\le \sum_{i=1}^\infty \big[1 - e^{-(1+\epsilon)p_i t}\big] + t\, 1_{A_\epsilon(t)^c} \\
&\le i_\epsilon + \sum_{i=i_\epsilon+1}^\infty \Big[1 - \exp\Big(-\frac{(1+\epsilon)^2\, c\, t}{i^\alpha}\Big)\Big] + t\, 1_{A_\epsilon(t)^c} \\
&\le i_\epsilon + \int_{i_\epsilon}^\infty \Big[1 - \exp\Big(-\frac{(1+\epsilon)^2\, c\, t}{x^\alpha}\Big)\Big]\,\mathrm{d}x + t\, 1_{A_\epsilon(t)^c} \quad \text{a.s.},
\end{aligned} \tag{3.5}$$

where the first inequality follows from $\overline{S}_1(t) \le N(t)$ a.s. and $\mathrm{E}(N(t) \mid \boldsymbol{q}(s), s \le t) = \mathrm{E}N(t) = t$, and the last inequality follows from that $1 - \exp\bigl(-(1+\epsilon)^2\, c\, t/i^\alpha\bigr)$ is decreasing in $i > 1$. For the integral in the last expression of (3.5), changing the variable to $u = (1+\epsilon)^2\, c\, t/x^\alpha$ and then integrating by parts, we have

$$\int_{i_\epsilon}^\infty \Bigl[1 - \exp\Bigl(-\frac{(1+\epsilon)^2\, c\, t}{x^\alpha}\Bigr)\Bigr]\, \mathrm{d}x = -i_\epsilon \Bigl[1 - \exp\Bigl(-\frac{(1+\epsilon)^2\, c\, t}{{i_\epsilon}^\alpha}\Bigr)\Bigr] + (1+\epsilon)^{2/\alpha}\, c^{1/\alpha}\, t^{1/\alpha} \int_0^{(1+\epsilon)^2 c t/{i_\epsilon}^\alpha} u^{-1/\alpha}\, e^{-u}\, \mathrm{d}u.$$

Therefore, after substituting this into (3.5), taking the expectation on the both sides of it, dividing by $t^{1/\alpha}$, then taking $t \to \infty$ and $\epsilon \downarrow 0$, we have under the condition that $\mathrm{P}(A_\epsilon(t)^c) = o(t^{-1+1/\alpha})$,

$$\limsup_{t\to\infty} \frac{\mathrm{E}\overline{S}_1(t)}{t^{1/\alpha}} \le c^{1/\alpha}\, \Gamma\Bigl(1 - \frac{1}{\alpha}\Bigr).$$

To verify (3.3), it remains to show the asymptotic lower bound. Similar to the above argument, we have from (3.4) that, under Assumption 1,

$$\begin{aligned}
\mathrm{E}(\overline{S}_1(t) \mid \boldsymbol{q}(s), s \le t) &\ge \sum_{i=1}^\infty [1 - e^{-(1-\epsilon)p_i t}]\, 1_{A_\epsilon(t)} \\
&\ge \sum_{i=i_\epsilon}^\infty \Bigl[1 - \exp\Bigl(-\frac{(1-\epsilon)^2\, c\, t}{i^\alpha}\Bigr)\Bigr]\, 1_{A_\epsilon(t)} \\
&\ge \int_{i_\epsilon}^\infty \Bigl[1 - \exp\Bigl(-\frac{(1-\epsilon)^2\, c\, t}{x^\alpha}\Bigr)\Bigr]\, \mathrm{d}x\, 1_{A_\epsilon(t)} \quad \text{a.s.}
\end{aligned} \tag{3.6}$$

Hence, similar to obtaining the asymptotic upper bound, we have

$$\liminf_{t\to\infty} \frac{\mathrm{E}\overline{S}_1(t)}{t^{1/\alpha}} \ge c^{1/\alpha}\, \Gamma\Bigl(1 - \frac{1}{\alpha}\Bigr),$$

since $\mathrm{P}(A_\epsilon(t)) \to 1$ as $t \to \infty$, so that (3.3) is obtained.

We next show that, under the condition of the lemma,

$$\mathrm{E}S_k(t) \sim \frac{c^{1/\alpha}}{\alpha\, k!}\, t^{1/\alpha}\, \Gamma\Bigl(k - \frac{1}{\alpha}\Bigr) \quad \text{as } t \to \infty, \quad k = 1, 2, \ldots. \tag{3.7}$$

By a similar argument to the above, we have from (3.1) and (3.2) that, under Assumption 1, for any $\epsilon \in (0,1)$,

$$\begin{aligned}
\mathrm{E}\bigl(S_k(t) \mid \boldsymbol{q}(s), s \le t\bigr) &\le \sum_{i=1}^\infty \frac{[(1+\epsilon)\, p_i\, t]^k}{k!}\, e^{-(1-\epsilon)p_i t} + t\, 1_{A_\epsilon(t)^c} \\
&\le i_\epsilon + \frac{1}{k!} \sum_{i=i_\epsilon+1}^\infty \Bigl[\frac{(1+\epsilon)^2\, c\, t}{i^\alpha}\Bigr]^k \exp\Bigl(-\frac{(1-\epsilon)^2\, c\, t}{i^\alpha}\Bigr) + t\, 1_{A_\epsilon(t)^c} \quad \text{a.s.}
\end{aligned} \tag{3.8}$$

Here, the summand of (3.8) is decreasing in sufficiently large i and we can choose $i_\epsilon > 0$ such that it is decreasing in $i \ge i_\epsilon$. Then,

$$\begin{aligned}
\mathrm{E}\bigl(S_k(t) \mid \boldsymbol{q}(s), s \le t\bigr) &\le i_\epsilon + \frac{1}{k!} \int_{i_\epsilon+1}^\infty \Bigl[\frac{(1+\epsilon)^2\, c\, t}{x^\alpha}\Bigr]^k \exp\Bigl(-\frac{(1-\epsilon)^2\, c\, t}{x^\alpha}\Bigr)\, \mathrm{d}x + t\, 1_{A_\epsilon(t)^c} \\
&= i_\epsilon + \frac{(1+\epsilon)^{2k}\, c^{1/\alpha}\, t^{1/\alpha}}{(1-\epsilon)^{2(k-1/\alpha)}\, \alpha\, k!} \int_0^{(1-\epsilon)^2 c t/{i_\epsilon}^\alpha} u^{k-1/\alpha-1}\, e^{-u}\, \mathrm{d}u + t\, 1_{A_\epsilon(t)^c} \quad \text{a.s.},
\end{aligned} \tag{3.9}$$

where the last equality follows from changing the variable to $u = (1-\epsilon)^2\, c\, t/x^\alpha$. Thus, taking the expectation on both sides of (3.9), dividing by $t^{1/\alpha}$, further taking $t \to \infty$ and $\epsilon \downarrow 0$, we have under the condition that $\mathrm{P}(A_\epsilon(t)^c) = o(t^{-1+1/\alpha})$,

$$\limsup_{t\to\infty} \frac{\mathrm{E}S_k(t)}{t^{1/\alpha}} \le \frac{c^{1/\alpha}}{\alpha\, k!}\, \Gamma\Big(k - \frac{1}{\alpha}\Big).$$

Similarly, we have under Assumption 1,

$$\begin{aligned}\mathrm{E}\big(S_k(t) \mid \boldsymbol{q}(s), s \le t\big) &\ge \sum_{i=1}^{\infty} \frac{[(1-\epsilon)\, p_i\, t]^k}{k!}\, e^{-(1+\epsilon)p_i t}\, 1_{A_\epsilon(t)} \\ &\ge \frac{1}{k!} \sum_{i=i_\epsilon}^{\infty} \Big[\frac{(1-\epsilon)^2\, c\, t}{i^\alpha}\Big]^k\, \exp\Big(-\frac{(1+\epsilon)^2\, c\, t}{i^\alpha}\Big)\, 1_{A_\epsilon(t)} \quad \text{a.s.},\end{aligned}$$

from which, a similar procedure to the above yields

$$\liminf_{t\to\infty} \frac{\mathrm{E}S_k(t)}{t^{1/\alpha}} \ge \frac{c^{1/\alpha}}{\alpha\, k!}\, \Gamma\Big(k - \frac{1}{\alpha}\Big),$$

since $\mathrm{P}(A_\epsilon(t)) \to 1$ as $t \to \infty$, and hence, (3.7) is obtained. Finally, (2.1) is immediate from (3.3) and (3.7). □

By Lemma 3.1 and the same argument as that after Proposition 2.1, we can derive the KRS distribution; that is, the right-hand side of (2.2), as the limit of $\mathrm{E}S_k(t)/\mathrm{E}\overline{S}_1(t)$ as $t \to \infty$ under the same condition as that of the lemma. Now, we show the almost sure convergence under a somewhat stronger condition.

Theorem 3.1 *Under Assumption 1, if* $\sum_{n=1}^{\infty} \mathrm{P}(A_\epsilon(n)^c) < \infty$, *then (2.2) holds for* $k = 1, 2, \ldots$.

To prove Theorem 3.1, we use the next lemma, where and thereafter, $\{N_i^\dagger(t)\}_{t\ge 0}$, $i = 1, 2, \ldots$, denotes a homogeneous Poisson process with intensity p_i and $\overline{S}_k^\dagger(t) = \sum_{i=1}^{\infty} 1_{\{N_i^\dagger(t)\ge k\}}$ for $k = 1, 2, \ldots$. Namely, $\{N_i^\dagger(t)\}_{t\ge 0}$ and $\overline{S}_k^\dagger(t)$ considered here are respectively nothing but $\{N_i(t)\}_{t\ge 0}$ and $\overline{S}_k(t)$ for the independent word occurrence model considered in the preceding section.

Lemma 3.2 *For any* $\epsilon > 0$, *there exists a* $\theta_\epsilon > 0$ *such that*

$$\mathrm{P}\big(\overline{S}_k(t) > (1+\epsilon)\, \mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\, t)\big) \le e^{-\theta_\epsilon \mathrm{E}\overline{S}_k^\dagger((1+\epsilon)t)} + \mathrm{P}(A_\epsilon(t)^c), \tag{3.10}$$

$$\mathrm{P}\big(\overline{S}_k(t) < (1-\epsilon)\, \mathrm{E}\overline{S}_k^\dagger((1-\epsilon)\, t)\big) \le e^{-\theta_\epsilon \mathrm{E}\overline{S}_k^\dagger((1-\epsilon)\, t)} + \mathrm{P}(A_\epsilon(t)^c). \tag{3.11}$$

Lemma 3.2 plays the role of (2.11) in the case of dependent word occurrences. (3.10) and (3.11) are also available when $m < \infty$ and are indeed exploited in the following subsections.

Proof: It is clear that

$$\mathrm{P}\big(\overline{S}_k(t) > (1+\epsilon)\, \mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\, t)\big) \le \mathrm{P}\big(\big\{\overline{S}_k(t) > (1+\epsilon)\, \mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\, t)\big\} \cap A_\epsilon(t)\big) + \mathrm{P}(A_\epsilon(t)^c),$$

and we consider the first term on the right-hand side above. Since a Poisson random variable is stochastically increasing in its mean, Theorem 1.A.14 in Shaked and Shanthikumar [8] implies that $N_i(t)\, 1_{\{Q_i(t)\le(1+\epsilon)p_i t\}} \le_{\mathrm{st}} N_i^\dagger((1+\epsilon)\, t)\, 1_{\{Q_i(t)\le(1+\epsilon)p_i t\}}$, $i = 1, 2, \ldots$, where "$\le_{\mathrm{st}}$" represents the usual stochastic order (see, e.g., Müller and Stoyan [5] or [8]). Thus, since $\overline{S}_k(t)$ is a.s. nondecreasing in $N_i(t)$, $i = 1, 2, \ldots$, we have $\overline{S}_k(t)\, 1_{A_\epsilon(t)} \le_{\mathrm{st}} \overline{S}_k^\dagger((1+\epsilon)\, t)\, 1_{A_\epsilon(t)}$, which implies that

$$\mathrm{P}\big(\big\{\overline{S}_k(t) > (1+\epsilon)\, \mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\, t)\big\} \cap A_\epsilon(t)\big) \le \mathrm{P}\big(\overline{S}_k^\dagger((1+\epsilon)\, t) > (1+\epsilon)\, \mathrm{E}\overline{S}_k^\dagger((1+\epsilon)\, t)\big).$$

Hence, the Chernoff-Hoeffding bound for the sum of 0-1 independent random variables yields (3.10). Inequality (3.11) is verified similarly. □

Proof of Theorem 3.1: By (3.10) in Lemma 3.2, we have

$$\mathrm{P}\Big(\frac{\overline{S}_k(t)}{\mathrm{E}\overline{S}_k^{\dagger}((1+\epsilon)\,t)} - 1 > \epsilon\Big) \le e^{-\theta_\epsilon \mathrm{E}\overline{S}_k^{\dagger}((1+\epsilon)t)} + \mathrm{P}\big(A_\epsilon(t)^c\big). \tag{3.12}$$

Here, since $\overline{S}_k^{\dagger}((1+\epsilon)\,t)$ is nothing but $\overline{S}_k((1+\epsilon)\,t)$ for the independent word occurrence model considered in the preceding section, Proposition 2.1 implies that $\mathrm{E}\overline{S}_k^{\dagger}((1+\epsilon)\,t) = \Theta(t^{1/\alpha})$ as $t \to \infty$, so that, under the condition of the theorem, (3.12) leads to

$$\sum_{n=1}^{\infty} \mathrm{P}\Big(\frac{\overline{S}_k(n)}{\mathrm{E}\overline{S}_k^{\dagger}((1+\epsilon)\,n)} - 1 > \epsilon\Big) < \infty.$$

Therefore, the Borel-Cantelli lemma implies that $\limsup_{n\to\infty} \overline{S}_k(n)/\mathrm{E}\overline{S}_k^{\dagger}((1+\epsilon)\,n) \le 1+\epsilon$ a.s. A similar argument based on (3.11) yields $\liminf_{n\to\infty} \overline{S}_k(n)/\mathrm{E}\overline{S}_k^{\dagger}((1-\epsilon)\,n) \ge 1-\epsilon$ a.s. Hence, applying Proposition 2.1 again and taking $\epsilon \downarrow 0$, we have $\overline{S}_k(n) \sim \mathrm{E}\overline{S}_k^{\dagger}(n)$ a.s. as $n \to \infty$ on $n \in \mathbb{N}$, which leads to (2.2) in the case where t goes to ∞ on $t \in \mathbb{N}$, but the extension to that on $t \in \mathbb{R}$ is easy. □

3.3 Case of $\alpha < 1$

We extend the results in Section 2.3. First, for $t \ge 0$ and $\epsilon > 0$, we define

$$A_\epsilon^{(m)}(t) = \Big\{\omega \in \Omega \;\Big|\; \max_{i\in\{1,\dots,m\}} \Big|\frac{Q_i^{(m)}(\omega,t)}{p_i^{(m)}\,t} - 1\Big| \le \epsilon\Big\}. \tag{3.13}$$

As in the independent word occurrence model, we first show the asymptotics of the expected size indices.

Lemma 3.3 *Under Assumption 2, if* $\mathrm{P}\big(A_\epsilon^{(m)}(\delta\,m/c)\big) \to 1$ *as* $m \to \infty$ *for some* $\delta > 0$*, then (2.3) holds for such* δ *and* $k = 1, 2, \dots$.

Note that, unlike the case of $\alpha > 1$, we here require that $\mathrm{P}\big(A_\epsilon^{(m)}(\delta\,m/c)^c\big)$ only vanishes as $m \to \infty$ but do not require the speed of vanishing for deriving the asymptotics of the expected size indices.

Proof: By (3.1), we have

$$\mathrm{E}\big(\overline{S}_k^{(m)} \mid \boldsymbol{q}^{(m)}(s), s \le t\big) = \sum_{i=1}^{m}\Big[1 - \sum_{l=0}^{k-1} \frac{Q_i^{(m)}(t)^l}{l!}\, e^{-Q_i^{(m)}(t)}\Big] \quad \text{a.s.}$$

For any $\epsilon > 0$, we have $(1-\epsilon)\,p_i^{(m)}\,t \le Q_i^{(m)}(t) \le (1+\epsilon)\,p_i^{(m)}\,t$ on $A_\epsilon^{(m)}(t)$ in (3.13), and thus, under Assumption 2,

$$\mathrm{E}\big(\overline{S}_k^{(m)} \mid \boldsymbol{q}^{(m)}(s), s \le t\big) \le m - \sum_{i=i_\epsilon}^{m}\sum_{l=0}^{k-1} \frac{1}{l!}\Big(\frac{(1-\epsilon)^2\,c\,t}{m^{1-\alpha}\,i^{\alpha}}\Big)^l \exp\Big(-\frac{(1+\epsilon)^2\,c\,t}{m^{1-\alpha}\,i^{\alpha}}\Big)\,1_{A_\epsilon^{(m)}(t)} \quad \text{a.s.,} \tag{3.14}$$

$$\mathrm{E}\big(\overline{S}_k^{(m)} \mid \boldsymbol{q}^{(m)}(s), s \le t\big) \ge (m - i_\epsilon)\,1_{A_\epsilon^{(m)}(t)} - \sum_{i=i_\epsilon}^{m}\sum_{l=0}^{k-1} \frac{1}{l!}\Big(\frac{(1+\epsilon)^2\,c\,t}{m^{1-\alpha}\,i^{\alpha}}\Big)^l \exp\Big(-\frac{(1-\epsilon)^2\,c\,t}{m^{1-\alpha}\,i^{\alpha}}\Big) \quad \text{a.s.} \tag{3.15}$$

Take the expectation on both sides of (3.14) and (3.15) and let $t = \delta\,m/c$ with $\delta > 0$. Then, since $\mathrm{P}\big(A_\epsilon^{(m)}(\delta\,m/c)\big) \to 1$ as $m \to \infty$ under the condition of the lemma, the remaining procedure is almost the same as that in the proof of Lemma 2.1. □

Next, we extend Theorem 2.1 and Corollary 2.1, where some vanishing speed of $\mathrm{P}\big(A_\epsilon^{(m)}(\delta\,m/c)^c\big) \to 0$ as $m \to \infty$ is required.

Theorem 3.2 *Under Assumption 2, if* $\sum_{m=1}^{\infty} \mathrm{P}(A_\epsilon^{(m)}(\delta\, m/c)^c) < \infty$ *for some* $\delta > 0$*, then (2.9) holds for such* δ *and* $k = 0, 1, 2, \ldots$*, and also (2.13) holds for* $k = 1, 2, \ldots$.

Proof: As in the proof of Theorem 3.1, let $\overline{S}_k^{\dagger(m)}(t)$, $k = 1, 2, \ldots$, just represent $\overline{S}_k^{(m)}(t)$ for the independent word occurrence model. Applying (3.10) in Lemma 3.2, we have

$$\mathrm{P}\Big(\frac{\overline{S}_k^{(m)}(\delta\, m/c)}{\mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\,\delta\, m/c)} - 1 > \epsilon\Big) \le e^{-\theta_\epsilon \mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\delta m/c)} + \mathrm{P}\big(A_\epsilon(\delta\, m/c)^c\big).$$

Here, (2.3) in Lemma 2.1 implies that $\mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\,\delta\, m/c) = \Theta(m)$ as $m \to \infty$, so that, we have under the condition of the theorem,

$$\sum_{m=1}^{\infty} \mathrm{P}\Big(\frac{\overline{S}_k^{(m)}(\delta\, m/c)}{\mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\,\delta\, m/c)} - 1 > \epsilon\Big) < \infty.$$

Therefore, the Borel-Cantelli lemma yields that $\limsup_{m\to\infty} \overline{S}_k^{(m)}(\delta\, m/c)/\mathrm{E}\overline{S}_k^{\dagger(m)}((1+\epsilon)\,\delta\, m/c) \le 1+\epsilon$ a.s. Similarly, by applying (3.11) in Lemma 3.2, we have $\liminf_{m\to\infty} \overline{S}_k^{(m)}(\delta\, m/c)/\mathrm{E}\overline{S}_k^{\dagger(m)}((1-\epsilon)\,\delta\, m/c) \ge 1-\epsilon$ a.s. Hence, applying Lemma 2.1 and taking $\epsilon \downarrow 0$, we have $\overline{S}_k^{(m)}(\delta\, m/c) \sim \mathrm{E}\overline{S}_k^{\dagger(m)}(\delta\, m/c)$ a.s. as $m \to \infty$, which implies (2.9) and (2.13). □

3.4 Case of $\alpha = 1$

Even in the case of $\alpha = 1$, a similar argument to the preceding subsections leads to the following results. The differences in the proofs are just that we apply Assumption 3 instead of Assumption 2 and replace $t = \delta\, m/c$ with $t = (\delta\, m\, \ln m)/c$, so that the proofs are omitted.

Lemma 3.4 *Under Assumption 3, if* $\mathrm{P}\big(A_\epsilon^{(m)}((\delta\, m\, \ln m)/c)\big) \to 1$ *as* $m \to \infty$ *for some* $\delta > 0$*, then (2.15) holds for such* δ *and* $k = 1, 2, \ldots$.

Theorem 3.3 *Under Assumption 3, if* $\sum_{m=1}^{\infty} \mathrm{P}\big(A_\epsilon^{(m)}((\delta\, m\, \ln m)/c)^c\big) < \infty$ *for some* $\delta > 0$*, then (2.18) holds for such* δ *and* $k = 0, 1, 2, \ldots$*, and also (2.19) holds for* $k = 1, 2, \ldots$.

4 Simulation experiments

We here validate the theoretical results discussed in the previous sections through simulation experiments; that is, we compare the limiting size index distributions derived in Sections 2 and 3 with the estimates of the empirical size index distributions by simulations. For the simulation experiments, the software R for statistical computing ([6]) was used. In each experiment, 100 independent replica of sample paths with length $t = 1000$ were executed, and the means and the 95% confidence intervals are displayed. Through the experiments, we will find not only that our analysis is valid but also that the derived limiting distributions well approximate the size index distributions even for relatively short (about 1000 words) texts.

4.1 Independent word occurrence model

Example 1 (Case of $\alpha > 1$) In the first example, we examine the independent word occurrence model with $\alpha > 1$. In the experiment, the word frequency distribution was given by $p_i = c/i^\alpha$, $i = 1, 2, \ldots$, and two cases of $\alpha = 1.1$ and $\alpha = 1.5$ were executed. The value of the normalization constant c was calculated as $c^{-1} = \sum_{i=1}^{i_0} 1/i^\alpha + {i_0}^{-\alpha+1}/(\alpha-1)$ with $i_0 = 2\times 10^7$, which is based on $\sum_{i>i_0} 1/i^\alpha \sim {i_0}^{-\alpha+1}/(\alpha-1)$ as

$i_0 \to \infty$, and the empirical size index distribution $S_k(t)/\overline{S}_1(t)$, $k = 1, 2, \ldots$, with $t = 1000$ was estimated by simulation. To compare the simulation estimates with the theoretical result, but to distinguish the two kinds of plots, we extend the range of the function on the right-hand side of (2.2) in Proposition 2.2 to $\mathbb{R}_+$; that is, applying $k! = \Gamma(k+1)$, the curves of the following function were evaluated;

$$\Psi(x;\alpha) = \frac{\Gamma(x - 1/\alpha)}{\alpha\,\Gamma(x+1)\,\Gamma(1-1/\alpha)}, \quad x > 0. \tag{4.1}$$

The value of $\Psi(x;\alpha)$ above, of course, coincides with that of the function on the right-hand side of (2.2) at $x = 1, 2, \ldots$. The result is displayed in Figure 1, where we can see good agreement of the theoretical result with the simulation estimates for both cases of $\alpha = 1.1$ and $\alpha = 1.5$.

Example 2 (Case of $\alpha < 1$) In the second example, we examine the independent word occurrence model with $\alpha < 1$. In the experiment, the word frequency distribution was given by $p_i^{(m)} = c_m/i^\alpha$, $i = 1, 2, \ldots, m$, with $\alpha = 0.7$. We recall that t and m are related by $t = \delta\, m/c$ in the result for $\alpha < 1$ in Section 2.3, and two cases of $m = 2\,t$ and $m = 10\,t$ with $t = 1000$ were executed. The value of the normalization constant c_m was calculated as ${c_m}^{-1} = \sum_{i=1}^m 1/i^\alpha$ and the empirical size index distribution $S_k^{(m)}(t)/\overline{S}_1^{(m)}(t)$, $k = 1, 2, \ldots$, was estimated by simulation. On the other hand, to evaluate the limiting size index distribution given on the right-hand side of (2.13) in Corollary 2.1, the value of constant δ was determined as follows. The value of c was first given by $c_m = c/m^{1-\alpha}$ from Assumption 2 and the value of δ was then determined by $\delta = c\,t/m = c_m\,t/m^\alpha$. As in the preceding example, we extend the range of the function on the right-hand side of (2.13) to $\mathbb{R}_+$, and applying the values of δ determined as above, the curves of the following function were evaluated;

$$\Psi(x;\alpha,\delta) = \frac{\delta^{1/\alpha}\,\Gamma(x - 1/\alpha,\,\delta)}{\alpha\,\Gamma(x+1)\left[\delta^{1/\alpha}\,\Gamma(1-1/\alpha,\,\delta) + 1 - e^{-\delta}\right]}, \quad x \geq 0. \tag{4.2}$$

The comparison result is displayed in Figure 2, where we can also see good agreement for both cases of $m = 2\,t$ and $m = 10\,t$. Furthermore, in the view of $\delta\big|_{m=10t} = c_{10t}\,t^{1-\alpha}/10^\alpha < c_{2t}\,t^{1-\alpha}/2^\alpha = \delta\big|_{m=2t}$, we can observe through the experimental result that the size index distribution tends to its degenerated form (2.14) as the value of δ decreases.

Example 3 (Case of $\alpha = 1$) The third example is almost the same as Example 2 but the case of $\alpha = 1$ is examined. As in the preceding example, two cases of $m = 2\,t$ and $m = 10\,t$ were executed with $t = 1000$. The value of the normalization constant c_m was calculated as ${c_m}^{-1} = \sum_{i=1}^m 1/i$ for the simulation runs. Then, to evaluate the theoretical formula (2.19) in Corollary 2.2, the value of δ was determined according to the relation $t = (\delta\, m\, \ln m)/c$; that is, the value of c was given by $c_m = c/\ln m$ from Assumption 3 and the value of δ was determined by $\delta = c\,t/(m\,\ln m) = c_m\,t/m$. The experimental result is displayed in Figure 3, where the curves of the function (4.2) with $\alpha = 1$ are plotted for the comparison with the plots of the simulation estimates. Again, we can see good agreement for both cases of $m = 2\,t$ and $m = 10\,t$.

4.2 Dependent word occurrence model

We here examine the dependent word occurrence model. To realize the dependence, we consider the random intensity process $\{\boldsymbol{q}(t)\}_{t\geq 0}$ governed by a two-state Markov chain $\{M(t)\}_{t\geq 0}$ on $\{0, 1\}$ such that

$$q_i^{(m)}(t) = \begin{cases} c_m^{(\mathrm{even})}/i^\alpha\, 1_{\{M(t)=0\}}, & i \text{ is even}, \\ c_m^{(\mathrm{odd})}/i^\alpha\, 1_{\{M(t)=1\}}, & i \text{ is odd}, \end{cases} \qquad t \geq 0; \tag{4.3}$$

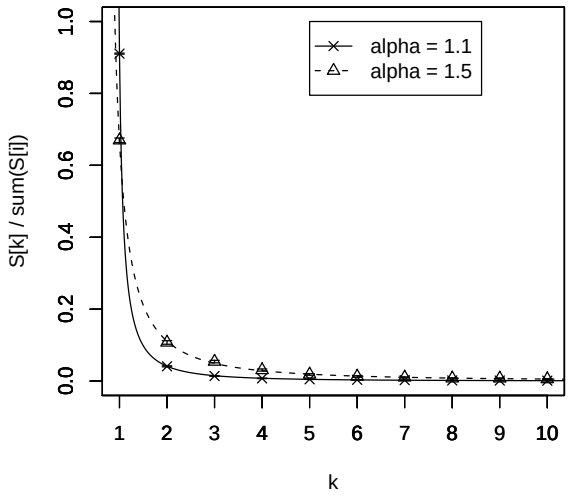

Figure 1: Experimental result for Example 1 (Independent word occurrence model with $\alpha > 1$)

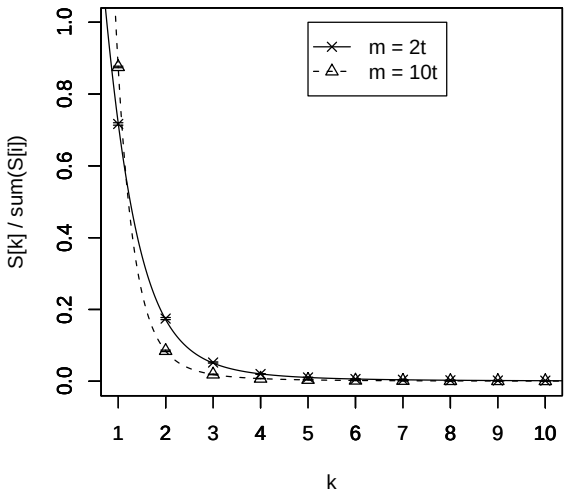

Figure 2: Experimental result for Example 2 (Independent word occurrence model with $\alpha = 0.7$)

that is, when the Markov chain is in state 0, only words with even indices are chosen and, when it is in state 1, only words with odd indices are chosen. In the experiments below, the parameter of sojourn times at each state was set at 1/5; that is, the Markov chain stays at a state during a time according to the exponential distribution with mean 5 and moves to the other state. The stationary distribution of this Markov chain is given as $(1/2, 1/2)$.

Example 4 (Case of $\alpha > 1$) As in Example 1, two cases of $\alpha = 1.1$ and $\alpha = 1.5$ were executed, where the values of $c^{(\mathrm{even})}$ and $c^{(\mathrm{odd})}$ in (4.3) were respectively calculated as ${c^{(\mathrm{even})}}^{-1} = \sum_{i=1}^{i_0} 1/(2\,i)^\alpha + (2\,i_0)^{-\alpha+1}/[2\,(\alpha-1)]$ and ${c^{(\mathrm{odd})}}^{-1} = \sum_{i=1}^{i_0} 1/(2\,i-1)^\alpha + (2\,i_0-1)^{-\alpha+1}/[2\,(\alpha-1)]$ with $i_0 = 10^7$. The experimental result is displayed in Figure 4, where the solid line and the dashed line are the same as those in Figure 1 since the right-hand side of (4.1) is irrelevant to the values of $c^{(\mathrm{even})}$ and $c^{(\mathrm{odd})}$. Even in the dependent word occurrence model, we can see the same feature as the independent word occurrence model; that is, good agreement of the theoretical result with the simulation experiment.

Example 5 (Case of $\alpha < 1$) As in Example 2, the parameter of the Zipf-type distribution was set at $\alpha = 0.7$, and two cases of $m = 2\,t$ and $m = 10\,t$ were executed for $t = 1000$. The values of $c_m^{(\mathrm{even})}$ and $c_m^{(\mathrm{odd})}$ in (4.3) were respectively calculated as ${c_m^{(\mathrm{even})}}^{-1} = \sum_{i=1}^{m/2} 1/(2\,i)^\alpha$ and ${c_m^{(\mathrm{odd})}}^{-1} = \sum_{i=1}^{m/2} 1/(2\,i-1)^\alpha$. To evaluate the theoretical formula (4.2), the value of constant δ was determined as follows. Since the stationary distribution of the Markov chain $\{M(t)\}_{t\geq 0}$ is given as $(1/2, 1/2)$, we have

$$p_i^{(m)} = \begin{cases} (1/2)\, c_m^{(\mathrm{even})}/i^\alpha, & i \text{ is even}, \\ (1/2)\, c_m^{(\mathrm{odd})}/i^\alpha, & i \text{ is odd}. \end{cases}$$

The value of c_m was set at the middle of $c_m^{(\mathrm{even})}/2$ and $c_m^{(\mathrm{odd})}/2$; that is, $c_m = (c_m^{(\mathrm{even})} + c_m^{(\mathrm{odd})})/4$, and the value of δ was then determined by $\delta = c_m\, t/m^\alpha$ as in Example 2. Applying the values of δ determined as such, the curves of the function on the right-hand side of (4.2) are plotted on Figure 5, as well as the simulation estimates. Even in this case, we can see good agreement for both cases of $m = 2\,t$ and $m = 10\,t$.

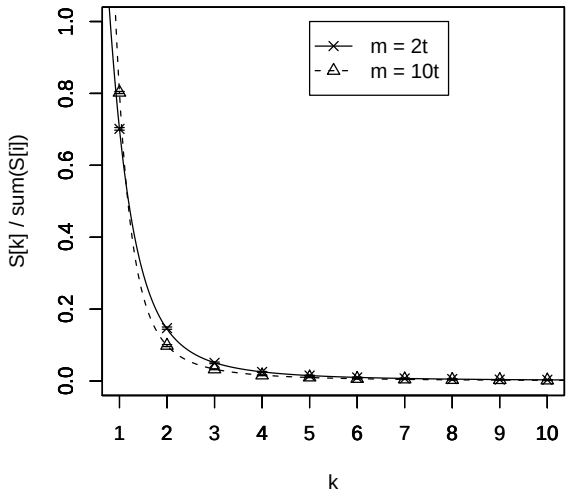

Figure 3: Experimental result for Example 3 (Independent word occurrence model with $\alpha = 1$)

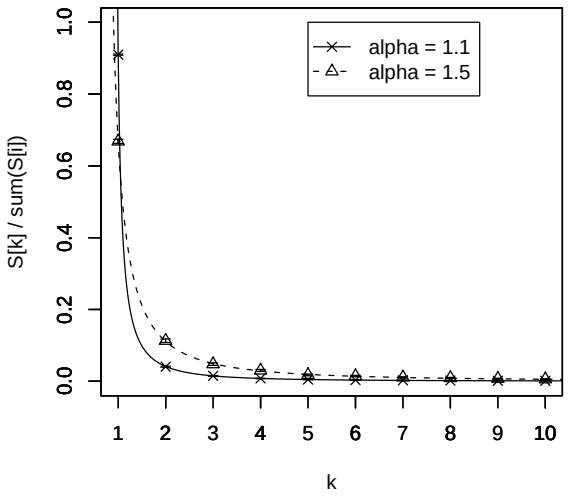

Figure 4: Experimental result for Example 4 (Dependent word occurrence model with $\alpha > 1$)

Example 6 (Case of $\alpha = 1$) Finally, we examine the dependent word occurrence model with $\alpha = 1$. The value of c_m was determined as the preceding example with $\alpha = 1$ and the value of δ was set at $\delta = c_m\, t/m$ as in Example 3. The experimental result is displayed in Figure 6, where we can also see good agreement.

Acknowledgments

The authors are grateful to Shigeru Mase for liberal encouragement and useful comments, as well as for some advice on the use of R. The first author wishes to thank Takafumi Kanamori for valuable discussions. Thanks are also due to Hirotaka Yamashita for some suggestion on the proof in Appendix.

References

[1] R. H. Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, 2001.

[2] P. J. Davis. Gamma function and related functions. In M. Abramowitz and C. A. Stegun, editors, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*, chapter 6, pages 253–293. Dover, 1972.

[3] S. Karlin. Central limit theorems for certain infinite urn schemes. *J. Math. Mech.*, 17:373–401, 1967.

[4] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[5] A. Müller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks*. John Wiley & Sons, 2002.

[6] R Development Core Team. *R: A Language and Environment for Statistical Computing*. `http://www.R-project.org`, 2007.

[7] A. Rouault. Lois de Zipf et sources markoviennes. *Ann. Inst. H. Poincaré Probab. Statist.*, 14:169–188, 1978.

[8] M. Shaked and J. G. Shanthikumar. *Stochastic Orders*. Springer-Verlag, 2007.

[9] M. Sibuya. Generalized hypergeometric, digamma and trigamma distributions. *Ann. Inst. Statist. Math.*, 31:373–390, 1979.

[10] M. Sibuya. A random clustering process. *Ann. Inst. Statist. Math.*, 45:459–465, 1993.

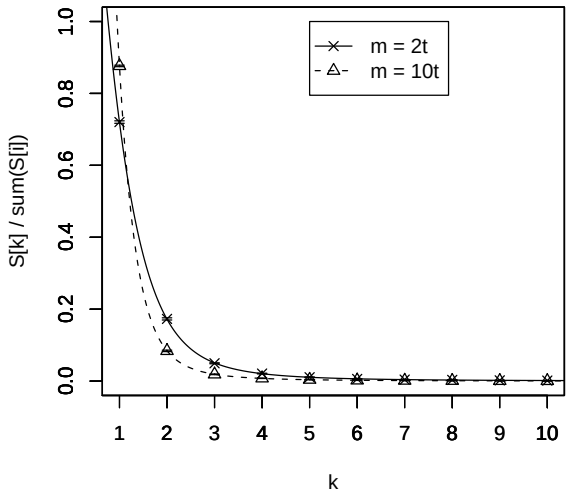

Figure 5: Experimental result for Example 5 (Dependent word occurrence model with $\alpha = 0.7$)

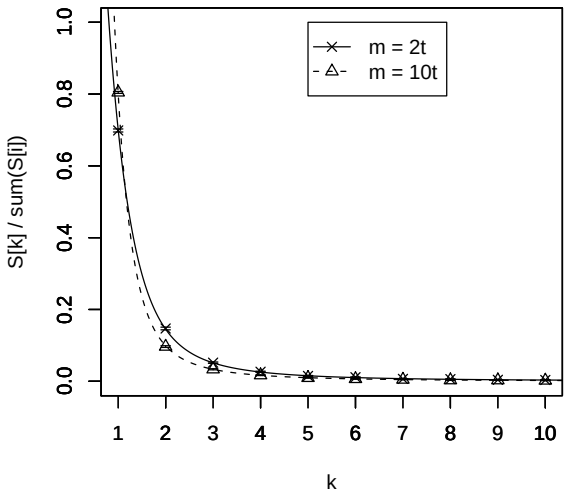

Figure 6: Experimental result for Example 6 (Dependent word occurrence model with $\alpha = 1$)

[11] I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Inform. Theory*, 37:1085–1094, 1991.

A Properties of $\Phi_k(\alpha, \delta)$ and $\Psi_k(\alpha, \delta)$

We here verify the assertions in Remarks 1 and 2 in Section 2. For $\alpha > 0$ and $\delta > 0$, $\Phi_k(\alpha, \delta)$, $k = 0, 1, 2, \ldots$, and $\Psi_k(\alpha, \delta)$, $k = 1, 2, \ldots$, are defined by

$$\Phi_k(\alpha, \delta) = \frac{\delta^{1/\alpha}}{\alpha\, k!}\, \Gamma\Big(k - \frac{1}{\alpha},\, \delta\Big), \quad k = 0, 1, 2, \ldots, \tag{A.1}$$

$$\Psi_k(\alpha, \delta) = \frac{\Phi_k(\alpha, \delta)}{\sum_{l=1}^{\infty} \Phi_l(\alpha, \delta)} = \frac{\Gamma\big(k - 1/\alpha,\, \delta\big)}{\alpha\, k! \left[\Gamma\big(1 - 1/\alpha,\, \delta\big) + \delta^{-1/\alpha}\,(1 - e^{-\delta})\right]}, \quad k = 1, 2, \ldots. \tag{A.2}$$

Lemma A.1 *(i) $\Phi_k(\alpha, \delta)$, $k = 0, 1, 2, \ldots$, in (A.1) gives a proper distribution on $\mathbb{Z}_+$ with $\Phi_k(\alpha, \delta) \sim (\delta^{1/\alpha}/\alpha)\, k^{-1-1/\alpha}$ as $k \to \infty$ for $\alpha > 0$ and $\delta > 0$. Furthermore, as $\delta \downarrow 0$, $\Phi_k(\alpha, \delta)$, $k = 0, 1, 2, \ldots$, degenerates as (2.10) for $\alpha > 0$.*

(ii) When $\alpha > 1$, $\Psi_k(\alpha, \delta)$, $k = 1, 2, \ldots$, in (A.2) reduces to the KRS distribution; that is, the right-hand side of (2.2), as $\delta \downarrow 0$, but when $\alpha \in (0, 1]$, it degenerates as (2.14).

Proof of (i): We first show that $\Phi_k(\alpha, \delta)$, $k = 0, 1, 2, \ldots$, gives a proper distribution when $1/\alpha \neq 1, 2, \ldots$. Applying the relation $\Gamma(x, y) = x^{-1}\left[\Gamma(x + 1, y) - y^x\, e^{-y}\right]$, $x \neq 0$, repeatedly, we have for $1/\alpha \neq 1, 2, \ldots$,

$$\sum_{l=0}^{k} \Phi_l(\alpha, \delta) = e^{-\delta} \sum_{l=0}^{k-1} \frac{\delta^l}{l!} - \frac{\delta^{1/\alpha}}{k!} \Big(k - \frac{1}{\alpha}\Big)\, \Gamma\Big(k - \frac{1}{\alpha},\, \delta\Big).$$

Thus, to check that $\sum_{k=0}^{\infty} \Phi_k(\alpha, \delta) = 1$, it suffices to show that the second term on the right-hand side above vanishes as $k \to \infty$. Note here that, for $k > 1/\alpha$,

$$0 \leq \frac{k - 1/\alpha}{k!}\, \Gamma\Big(k - \frac{1}{\alpha},\, \delta\Big) \leq \frac{k - 1/\alpha}{k!}\, \Gamma\Big(k - \frac{1}{\alpha}\Big).$$

Thus, applying $\Gamma(x+1) = x\,\Gamma(x)$ and $k!\,k^x/[x\,(x+1)\cdots(x+k)] \to \Gamma(x)$ as $k \to \infty$ for $x \neq 0, -1, -2, \ldots$ (see, e.g., [2]), we have

$$\frac{k-1/\alpha}{k!}\,\Gamma\Big(k-\frac{1}{\alpha}\Big) = \frac{(k-1/\alpha)\,(k-1-1/\alpha)\cdots(-1/\alpha)}{k!}\,\Gamma\Big(-\frac{1}{\alpha}\Big) \sim k^{-1/\alpha} \to 0 \quad \text{as } k \to \infty. \tag{A.3}$$

On the other hand, when $1/\alpha = 1, 2, \ldots$, we use the following relations which are confirmed inductively;

$$\Gamma(n, y) = e^{-y}\,(n-1)!\sum_{k=0}^{n-1}\frac{y^k}{k!}, \quad n = 1, 2, \ldots, \tag{A.4}$$

$$\Gamma(-n, y) = \frac{e^{-y}}{n!}\sum_{k=1}^{n}(-1)^{n-k}\,(k-1)!\,y^{-k} + \frac{(-1)^n}{n!}\,\Gamma(0, y), \quad n = 0, 1, 2, \ldots. \tag{A.5}$$

Letting $1/\alpha = n \in \mathbb{N}$, we split the sum of $\Phi_k(\alpha, \delta)$ in (A.1) over $k = 0, 1, 2, \ldots$ into two terms;

$$\sum_{k=0}^{\infty}\Phi_k(1/n, \delta) = \sum_{k=0}^{n}\frac{n\,\delta^n}{k!}\,\Gamma(-(n-k), \delta) + \sum_{k=n+1}^{\infty}\frac{n\,\delta^n}{k!}\,\Gamma(k-n, \delta). \tag{A.6}$$

We first consider the second term on the right-hand side above. Applying (A.4), we have

$$\begin{aligned}\sum_{k=n+1}^{\infty}\frac{n\,\delta^n}{k!}\,\Gamma(k-n, \delta) &= n\,e^{-\delta}\sum_{k=n+1}^{\infty}\frac{(k-n-1)!}{k!}\sum_{l=0}^{k-n-1}\frac{\delta^{n+l}}{l!}\\ &= n\,e^{-\delta}\sum_{l=0}^{\infty}\frac{\delta^{n+l}}{l!}\sum_{k=n+l+1}^{\infty}\frac{(k-n-1)!}{k!}\\ &= e^{-\delta}\sum_{l=0}^{\infty}\frac{\delta^{n+l}}{(n+l)!},\end{aligned} \tag{A.7}$$

where the third equality follows since

$$\begin{aligned}\sum_{k=n+l+1}^{\infty}\frac{(k-n-1)!}{k!} &= \frac{1}{n}\sum_{k=n+l+1}^{\infty}\Big[\frac{1}{(k-1)\cdots(k-n)} - \frac{1}{k\cdots(k-n+1)}\Big]\\ &= \frac{1}{n}\,\frac{1}{(n+l)\cdots(l+1)}.\end{aligned}$$

Now, we consider the first term on the right-hand side of (A.6). Applying (A.5), we have

$$\sum_{k=0}^{n}\frac{n\,\delta^n}{k!}\,\Gamma(-(n-k), \delta) = \sum_{k=0}^{n}\frac{n\,e^{-\delta}}{k!\,(n-k)!}\sum_{l=1}^{n-k}(-1)^{n-k-l}\,(l-1)!\,\delta^{n-l} + n\,\delta^n\,\Gamma(0, \delta)\sum_{k=0}^{n}\frac{(-1)^{n-k}}{k!\,(n-k)!},$$

and the second term on the right-hand side above is clearly zero by the binomial theorem. For the first term above, we have

$$\begin{aligned}\sum_{k=0}^{n}\frac{n\,e^{-\delta}}{k!\,(n-k)!}\sum_{l=1}^{n-k}(-1)^{n-k-l}\,(l-1)!\,\delta^{n-l} &= n\,e^{-\delta}\sum_{l=1}^{n}(l-1)!\,\delta^{n-l}\sum_{k=0}^{n-l}\frac{(-1)^{n-l-k}}{k!\,(n-k)!}\\ &= e^{-\delta}\sum_{l=1}^{n}\frac{\delta^{n-l}}{(n-l)!}\\ &= e^{-\delta}\sum_{l=0}^{n-1}\frac{\delta^l}{l!},\end{aligned} \tag{A.8}$$

where, in the second equality, we use $\sum_{k=0}^{l}\binom{n}{k}(-1)^k = \binom{n-1}{l}(-1)^l$, $l = 0, 1, \ldots, n-1$, which is confirmed by the observation that, when the coefficient of x^k in $(1-x)^{n-1}$ is denoted by a_k, $k =$

$0, \ldots, n-1$, then the coefficient of x^k in $(1-x)^n$ is given by $b_k = a_k - a_{k-1}$ for $k = 1, \ldots, n-1$ and $b_0 = a_0 = 1$ since $(1-x)^n = (1-x)^{n-1} - x\,(1-x)^{n-1}$. Hence, substituting (A.7) and (A.8) into (A.6), we have $\sum_{k=0}^{\infty} \Phi(1/n, \delta) = 1$ for $n = 1, 2, \ldots$.

Next, we confirm the asymptotics of $\Phi_k(\alpha, \delta)$ as $k \to \infty$. By (A.1), we have for $k > 1/\alpha$,

$$\Phi_k(\alpha, \delta) = \frac{\delta^{1/\alpha}}{\alpha\, k!}\, \Gamma\Big(k - \frac{1}{\alpha}\Big) - \frac{\delta^{1/\alpha}}{\alpha\, k!} \int_0^{\delta} e^{-u}\, u^{k-1/\alpha-1}\, \mathrm{d}u. \tag{A.9}$$

The integrand above takes the maximum at $u = k - 1/\alpha - 1$, which is greater than δ for sufficiently large k. Thus, for large k,

$$\frac{\delta^{1/\alpha}}{\alpha\, k!} \int_0^{\delta} e^{-u}\, u^{k-1/\alpha-1}\, \mathrm{d}u \le \frac{e^{-\delta}\, \delta^k}{\alpha\, k!} \sim \frac{e^{-\delta}}{\sqrt{2\,\pi}\,\alpha}\, \frac{(\delta\, e)^k}{k^{k+1/2}} = o(k^{-1-1/\alpha}) \quad \text{as } k \to \infty,$$

where "$\sim$" follows from Stirling's formula $k! \sim \sqrt{2\,\pi}\, k^{k+1/2}\, e^{-k}$ as $k \to \infty$. On the other hand, for the first term on the right-hand side of (A.9), if $1/\alpha \neq 1, 2, \ldots$, similar to obtaining (A.3), we have

$$\frac{\delta^{1/\alpha}}{\alpha\, k!}\, \Gamma\Big(k - \frac{1}{\alpha}\Big) = \frac{\delta^{1/\alpha}}{\alpha}\, \frac{k - 1/\alpha - 1) \cdots (-1/\alpha)}{k!}\, \Gamma\Big(-\frac{1}{\alpha}\Big) \sim \frac{\delta^{1/\alpha}}{\alpha}\, k^{-1-1/\alpha} \quad \text{as } k \to \infty.$$

Also, if $1/\alpha = 1, 2, \ldots$, then

$$\frac{\delta^{1/\alpha}}{\alpha\, k!}\, \Gamma\Big(k - \frac{1}{\alpha}\Big) = \frac{\delta^{1/\alpha}}{\alpha}\, \frac{(k - 1/\alpha - 1)!}{k!} \sim \frac{\delta^{1/\alpha}}{\alpha}\, k^{-1-1/\alpha} \quad \text{as } k \to \infty.$$

Now, we show the limits of $\Phi_k(\alpha, \delta)$, $k = 0, 1, 2, \ldots$, as $\delta \downarrow 0$. Since $\Gamma(x, y) \to \Gamma(x) < \infty$ as $y \downarrow 0$ for $x > 0$, it is clear from (A.1) that $\Phi_k(\alpha, \delta) \to 0$ as $\delta \downarrow 0$ for $k > 1/\alpha$. For $k \le 1/\alpha$, since $\Gamma(x, y) \to \infty$ as $y \downarrow 0$ for $x \le 0$, applying de l'Hospital's rule, we have $\Phi_k(\alpha, \delta) \sim e^{-\delta}\, \delta^k / k!$ as $\delta \downarrow 0$, so that (2.10) is derived. □

Proof of (ii): We investigate the limiting property of $\Psi(\alpha, \delta)$, $k = 1, 2, \ldots$, as $\delta \downarrow 0$. Applying de l'Hospital's rule to the second term in the brackets of the denominator in (A.2), we have

$$\delta^{-1/\alpha}\, (1 - e^{-\delta}) \sim \frac{\alpha\, e^{-\delta}}{\delta^{1/\alpha - 1}} \to \begin{cases} 0, & \alpha > 1, \\ 1, & \alpha = 1, \\ +\infty, & 0 < \alpha < 1, \end{cases} \qquad \text{as } \delta \downarrow 0. \tag{A.10}$$

Thus, since $\Gamma(x, y) \to \Gamma(x) < \infty$ as $y \downarrow 0$ for $x > 0$, the limit of $\Psi_k(\alpha, \delta)$ as $\delta \downarrow 0$ clearly reduces to the KRS distribution when $\alpha > 1$. When $\alpha = 1$, since $\Gamma(0, y) \to \infty$ as $y \downarrow 0$, it is also clear from (A.2) and (A.10) that (2.14) holds. Consider the case of $0 < \alpha < 1$. Note that $\Gamma(x, y) \to \Gamma(x) < \infty$ for $x > 0$ and $\Gamma(x, y) \to \infty$ for $x \le 0$ as $y \downarrow 0$. Thus, when $k > 1/\alpha$, clearly $\Psi_k(\alpha, \delta) \to 0$ as $\delta \downarrow 0$ by (A.2) and (A.10). To consider the case of $k \le 1/\alpha$, we deform the right-hand side of (A.2) as

$$\Psi_k(\alpha, \delta) = \frac{1}{\alpha\, k!}\, \frac{\Gamma(k - 1/\alpha, \delta)}{\Gamma(1 - 1/\alpha, \delta)} \left[1 + \frac{\delta^{-1/\alpha}\, (1 - e^{-\delta})}{\Gamma(1 - 1/\alpha, \delta)}\right]^{-1}.$$

Here, applying de l'Hospital's rule to the second term in the brackets, we have

$$\frac{\delta^{-1/\alpha}\, (1 - e^{-\delta})}{\Gamma(1 - 1/\alpha, \delta)} \sim \frac{(1/\alpha)\, \delta^{-1}\, (1 - e^{-\delta}) - e^{-\delta}}{e^{-\delta}} \to \frac{1}{\alpha} - 1 \quad \text{as } \delta \downarrow 0,$$

where we use (A.10). Furthermore, applying de l'Hospital's rule again, we have $\Gamma(k - 1/\alpha, \delta)/\Gamma(1 - 1/\alpha, \delta) \sim \delta^{k-1}$ as $\delta \downarrow 0$. Hence, we eventually have (2.14) for $0 < \alpha < 1$. □