Statistics and Probability Letters 113 (2016) 49–53

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Bin sizes in time-inhomogeneous infinite Polya processes

Dudley Stark

School of Mathematical Sciences, Queen Mary, University of London, London E1 4NS, United Kingdom

ARTICLE INFO

Article history:
Received 21 November 2015
Received in revised form 12 February 2016
Accepted 17 February 2016
Available online 4 March 2016

Keywords:
Polya model
Hoppe model
Preferential attachment

ABSTRACT

At the nth step of a time-inhomogeneous infinite Polya process, either a new bin is created and a ball is put in that bin, or a ball is put into an existing bin, in which case the bin is chosen according to a preferential attachment type rule. We introduce a new class of such processes and find the asymptotics of the expected number of bins of size k, for k fixed, as $n \to \infty$.

1. Introduction

Simon (1955) introduced a model for generating observed word frequencies in data sets from English and other languages. Initially, at time $n = 1$, there is one word and it is assumed that one word is added at time n for $n \geq 2$. At time n, let $N(n)$ denote the number of different words and let $W_i(n)$ denote the number of copies of word i, so that $W_1(n) + W_2(n) + \cdots + W_{N(n)} = n$. At each time step, either a new word is added with probability α, or a copy of an existing word is added with probability $1 - \alpha$. If a copy of an existing word is added at time $n + 1$, then a copy of word i is made with probability $\frac{W_i(n)}{n}$. The way existing words are chosen to be copied is similar to the preferential attachment rule used in generating scale-free graphs; see Albert and Barabási (2002) and Barabási and Albert (1999). Simon's model has a limiting distribution which was first obtained in a paper of Yule (1925) for a different preferential attachment model. It is also related to the Polya urn model (Johnson and Kotz, 1977) and is equivalent to the infinite Polya process model studied in Chung and Lu (2006). In the infinite Polya process, words are replaced by bins and the number of copies of a given word corresponds to the number of balls in a bin. Initially, there is one bin containing one ball, and whenever a new bin is created, it is created containing one ball.

Another widely studied urn model is that of Hoppe (1984). Its description is identical to Simon's model, except that at the nth time step (from n to $n+1$ balls) a new bin (containing one ball) is added with probability $\alpha_n = \theta/(\theta+n)$ for a parameter $\theta > 0$. Hoppe's urn model was extended by Dubins and Pitman, announced in Aldous and Hennequin (1985), to a way of constructing random permutations called the Chinese restaurant process. The distribution of the process of cycle counts in the Chinese restaurant process is called the Ewens sampling formula and is of fundamental importance in population genetics (Ewens, 2004).

Simon (1955) was interested in altering his model so that α is a function of n and heuristically discussed two examples with varying α_n. It was noted by Eriksson and Sjöstrand (2012) that nothing is known for dependencies of α_n on n beyond the Polya and Hoppe models. We are motivated, therefore, to study the infinite Polya process under the condition

$$\alpha_n = \theta n^{r-1} + O\left(n^{r-1-\epsilon}\right), \quad n \geq 1, \tag{1}$$

E-mail address: d.s.stark@qmul.ac.uk.

http://dx.doi.org/10.1016/j.spl.2016.02.010
0167-7152/

where $r \leq 1$ and $\epsilon > 0$ are parameters. If $r = 1$, then $\theta \in [0, 1]$ must hold because the α_n are probabilities, but otherwise we allow $\theta \geq 0$. The case $r = 0$ includes Hoppe's model and $r = 1$ includes Simon's model. The cases $r \in (-\infty, 0) \cup (0, 1)$ do not previously seem to have been studied. The cases $r \in (0, 1)$ interpolate between the two well known models already described. The $O(\cdot)$ term in (1) introduces generality into our model.

Let $W(n, k)$ be the number of bins containing k balls and let $f(n, k) = \mathbf{E}(W(n, k))$. We find asymptotics for $f(n, k)$ for the infinite Polya process with α_n given by (1). We assume $W(0, 1) = 1$ and $W(0, k) = 0$ for $k > 1$. Our main result is

Theorem 1. *For the infinite Polya process with α_n given by (1) with $r = 1$, for fixed $k \geq 1$,*

$$f(n,k) = M_k n + O\left(n^{1-\epsilon'}\right) \tag{2}$$

for a constant $\epsilon' > 0$, where M_k is given recursively by

$$M_1 = \frac{\theta}{2-\theta}, \qquad M_k = M_{k-1}\frac{(1-\theta)(k-1)}{1+k(1-\theta)}, \quad k \geq 2.$$

For α_n given by (1) with $r \in (-1, 1)$, for fixed $k \geq 1$,

$$f(n,k) = \frac{\theta(k-1)!}{(r+1)^{(k)}} n^r + O\left(n^{r-\epsilon'}\right) \tag{3}$$

for a constant $\epsilon' > 0$, where $x^{(k)} = x(x+1)(x+2)\cdots(x+k-1)$ is notation for rising factorial.

Theorem 1 gives a bound on $f(n, k)$ when $r \leq -1$: taking $\theta = 0$ in the theorem shows that if $\alpha_n = O\left(n^{r-1-\epsilon}\right)$ for some $r \in (-1, 1)$, then $f(n, k) = O\left(n^{r-\epsilon'}\right)$. It should be possible to extend the range of r in the theorem to $r \leq -1$ by our methods, in which case the formulae for the $f(n, k)$ may involve logarithmic factors of n. For example, if $r = -1$, $f(n, 1) = \frac{\theta \log n}{n}(1 + O(n^{-\epsilon'}))$.

The following corollary examines $\lim_{n\to\infty} f(n, k)/n^r$. In the cases other than $\theta = 0$ and $r = 1, \theta = 1$, the limit is scale-free as $k \to \infty$, meaning it decays polynomially.

Corollary 1. *For all r, if $\theta = 0$, then $f(n, k) = o(n^r)$ for all k as $n \to \infty$. If $r = 1, \theta = 1$, we have*

$$\lim_{n\to\infty} f(n,k)/n = \begin{cases} 1 & \text{if } k = 1; \\ 0 & \text{if } k > 1. \end{cases}$$

As $k \to \infty$, we have

$$\lim_{n\to\infty} f(n,k)/n^r \sim \begin{cases} \dfrac{\theta}{2-\theta}\Gamma\left(\dfrac{3-2\theta}{1-\theta}\right)k^{-\frac{2-\theta}{1-\theta}} & \text{if } r = 1,\ \theta \in (0,1); \\ \theta\Gamma(r+1)k^{-r-1} & \text{if } r \in (-1,1),\ \theta > 0. \end{cases}$$

Proof. If $\theta = 0$, then $M_k = 0$ for all $k \geq 1$ and $f(n, k) = O\left(n^{r-\epsilon'}\right) = o(n^r)$. If $\theta = 1$, then $M_1 = 1$ and $M_k = 0$ for $k \geq 2$. Calculating as in Chung and Lu (2006) and Simon (1955), for $k \geq 2$ we have

$$M_k = M_{k-1}\frac{k-1}{k+\frac{1}{1-\theta}}$$

and so

$$M_k = \frac{\theta}{2-\theta}\prod_{\ell=2}^{k}\frac{\ell-1}{\ell+\frac{1}{1-\theta}} = \frac{\theta}{2-\theta}\frac{\Gamma(k)\Gamma\left(2+\frac{1}{1-\theta}\right)}{\Gamma\left(k+1+\frac{1}{1-\theta}\right)} = \frac{\theta}{2-\theta}\frac{\Gamma(k)\Gamma\left(\frac{3-2\theta}{1-\theta}\right)}{\Gamma\left(k+\frac{2-\theta}{1-\theta}\right)}.$$

The well-known property of the Gamma function

$$\frac{\Gamma(k)}{\Gamma(k+\rho)} \sim k^{-\rho}, \quad k \to \infty,$$

results in

$$M_k \sim \frac{\theta}{2-\theta}\Gamma\left(\frac{3-2\theta}{1-\theta}\right)k^{-\frac{2-\theta}{1-\theta}},$$

giving us the result for $r = 1, \theta \in (0, 1)$. We also have

$$\frac{\theta(k-1)!}{(r+1)^{(k)}} = \frac{\theta\Gamma(k)\Gamma(r+1)}{\Gamma(r+k+1)} \sim \theta\Gamma(r+1)k^{-r-1}$$

giving the result for $r \in (-1, 1), \theta > 0$. ∎