Bin sizes in time-inhomogeneous infinite Polya processes

Dudley Stark

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

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 $n \to \infty$.

Keywords: Polya model, Hoppe model, preferential attachment

1. Introduction

Simon [12] introduced the following 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 each time step, either a new word is added with probability α, or a copy of an existing word is added with probability $1 - \alpha$. Suppose that a copy of an existing word is added at time $n + 1$ and let $W(n, k)$ be the number of words with exactly k copies at time n. Let $f(n, k) = \mathbf{E}(W(n, k))$. Simon assumed either

- that the probability that the $(n+1)$-st word is a word that has already appeared k times is proportional to $kf(n, k)$.
- that the probability a *particular* word occur next be proportional to the number of its previous occurrences.

For our purposes, either of these assumptions can be used, the reason being that in either case the probability that with probability proportional to $kf(n, k)$ a word appearing k times is copied. The way existing words are chosen to be copied is similar to the preferential attachment rule used in generating scale-free graphs; see [1, 4].

Simon's model is related to the Polya urn model [11] and is basically equivalent to the infinite Polya process model studied in Chung and Lu [5]. 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.

Another widely studied urn model is that of Hoppe [10]. 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 [2], to a way of constructing random permutations called the Chinese restaurant process. The

Email address: d.s.stark@qmul.ac.uk (Dudley Stark)

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 [8].

Simon [12] 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 [6] 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

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

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

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

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 ion $f(n, k)$ 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 $f(n, k) = o\left(n^r\right)$ 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

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

Proof If $\theta = 0$, then $M_k = 0$ for all $k \geq 1$ and $f(n,k) = O\left(n^{r-\epsilon'}\right) = o\left(n^r\right)$. If $\theta = 1$, then $M_1 = 1$ and $M_k = 0$ for $k \geq 2$. Calculating as in [5, 12], 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 $r \in (-1,1)$, $\theta > 0$. ∎

For the time-homogeneous Polya process, (2) was obtained previously in [5, 12]. Watterson [13] found falling factorial moments for the Ewens sampling formula implying (3) for the Hoppe urn model, in which case $r = 0$ and $f(n,k)$ converges to θ/k. It also holds that the distribution of $W(n,k)$ converges in total variation distance to Poisson(θ/k). A great deal more information is contained in [3] on the process $(W(n,1), W(n,2), \ldots)$ for the Ewens sampling formula and related random combinatorial structures.

Limit shapes are an area of interest connected with this research; see [6, 7]. Theorem 1 does not give results regarding a possible limit shape, the reason being that we have not obtained information on the largest bin size, but only on the smallest k bin sizes for a fixed k. The probability that there is a bin size larger than k is bounded below by the probability that in the first $k+1$ steps of the process no new urns are created and that all of the balls are put into the first urn, which is $(1-\theta)^{k+1} > 0$. Erlihson and Granovsky [7] prove a theorem showing that the existence of a limiting shape implies the property that there is no gelation, which in our context has to do with the almost sure behaviour of the size of the bin with the most number of balls.

The framework of our proof of Theorem 1 is introduced in Section 2 and Theorem 1 is proved in Section 3.

2. Preliminaries

The starting point for our analysis are recursions for $f(n,k)$ which were also used in [5, 12]. There is a slight generalisation here, because we let α_n vary with n. Let $\mathcal{F}_n$ be the σ-algebra generated by $\{W(j,k),\ j \le n, k \ge 1\}$. By the description of the process in Section 1, we have

$$\begin{aligned} f(n+1,1) &= \mathbf{E}(W(n+1,1) \\ &= \mathbf{E}(\mathbf{E}(W(n+1,1)|\mathcal{F}_n)) \\ &= \mathbf{E}\left(W(n,1) + \alpha_n - (1-\alpha_n)\frac{W(n,1)}{n}\right) \\ &= f(n,1) + \alpha_n - \frac{(1-\alpha_n)f(n,1)}{n} \\ &= \left(1 - \frac{1-\alpha_n}{n}\right) f(n,1) + \alpha_n \end{aligned} \tag{4}$$

and, for $k \ge 2$,

$$\begin{aligned} f(n+1,k) &= \mathbf{E}(W(n+1,k) \\ &= \mathbf{E}(\mathbf{E}(W(n+1,k)|\mathcal{F}_n)) \\ &= \mathbf{E}\left(W(n,k) + (1-\alpha_n)\frac{(k-1)W(n,k-1)}{n} - (1-\alpha_n)\frac{kW(n,k)}{n}\right) \\ &= f(n,k) + (1-\alpha_n)\frac{(k-1)f(n,k-1)}{n} - (1-\alpha_n)\frac{kf(n,k)}{n} \\ &= \left(1 - \frac{(1-\alpha_n)k}{n}\right) f(n,k) + \frac{(1-\alpha_n)(k-1)f(n,k-1)}{n}. \end{aligned} \tag{5}$$

Both (4) and (5) are recursions of the form

$$c_{n+1} = a_n c_n + b_n, \quad n \ge 1. \tag{6}$$

It is easy to verify by induction that the solution to recursion (6) is

$$c_n = \sum_{j=1}^{n-1} b_j \prod_{\ell=j+1}^{n-1} a_\ell + c_1 \prod_{j=1}^{n-1} a_j. \tag{7}$$

We will use (7) to find expressions for the $f(n,k)$ and then find the asymptotics of the expressions inductively. In doing so, we will need the following result. In proving the lemma and in proofs in Section 3, we will use the asymptotic for the harmonic numbers (see [9]) $\sum_{\ell=1}^{n} 1/\ell = \log n + \gamma + O(n^{-1})$, where γ is Euler's constant and the fact that $\sum_{j=2}^{n} j^\beta = \frac{n^{\beta+1}}{\beta+1} + O(n^\beta)$ for any constant $\beta > -1$. We let $\epsilon' > 0$ denote a constant which can change from line to line.

Lemma 1. *Let $k \ge 1$ be an integer. As $n \to \infty$, for all integers $j \in [k, n-1]$, the following estimate holds,*

where the $\epsilon' > 0$ are constants and where the constants in the $O(\cdot)$ bounds are universal.

$$\prod_{\ell=j+1}^{n-1}\left(1-\frac{(1-\alpha_\ell)k}{\ell}\right)=\begin{cases}(j/n)^{k-k\theta}\left(1+O\left(j^{-\epsilon'}\right)\right) & \text{if } r=1;\\ (j/n)^{k}\left(1+O\left(j^{-\epsilon'}\right)\right) & \text{if } 0\le r<1.\end{cases}$$

Proof For all $\ell \ge k+1$, $0 \le (1-\alpha_\ell)k/\ell \le k/(k+1) < 1$. Thus, for all $j \in [k,n]$ we have

$$\begin{aligned}\prod_{\ell=j+1}^{n-1}\left(1-\frac{(1-\alpha_\ell)k}{\ell}\right) &= \exp\left(-\sum_{\ell=j+1}^{n-1}\frac{(1-\alpha_\ell)k}{\ell}+O\left(\sum_{\ell=j+1}^{n-1}\frac{1}{\ell^2}\right)\right)\\ &= \exp\left(-\sum_{\ell=j+1}^{n-1}\frac{k}{\ell}+\sum_{\ell=j+1}^{n-1}k\left[\theta\ell^{r-2}+O\left(\ell^{r-2-\epsilon}\right)\right]+O\left(j^{-1}\right)\right)\\ &= \exp\left(-k\log\left(\frac{n}{j}\right)+\sum_{\ell=j+1}^{n-1}k\left[\theta\ell^{r-2}+O\left(\ell^{r-2-\epsilon}\right)\right]+O\left(j^{-1}\right)\right). \qquad (8)\end{aligned}$$

If $r=1$, then (8) becomes

$$\prod_{\ell=j+1}^{n-1}\left(1-\frac{(1-\alpha_\ell)k}{\ell}\right)=\left(\frac{j}{n}\right)^{k-k\theta}\exp\left(O\left(j^{-\epsilon'}\right)\right)=\left(\frac{j}{n}\right)^{k-k\theta}\left(1+O\left(j^{-\epsilon'}\right)\right),$$

for some $0<\epsilon'\le 1$. If $r\in(-1,1)$, then

$$\prod_{\ell=j+1}^{n-1}\left(1-\frac{(1-\alpha_\ell)k}{\ell}\right)=\left(\frac{j}{n}\right)^{k}\exp\left(O\left(j^{-\epsilon'}\right)\right)=\left(\frac{j}{n}\right)^{k}\left(1+O\left(j^{-\epsilon'}\right)\right).$$

■

3. Proofs

We will use Lemma 1 and induction on k to prove Theorem 1.
The identities (4) and (7) along with $f(n,1)=1$, imply

$$\begin{aligned}f(n,1) &= \sum_{j=1}^{n-1}\alpha_j\prod_{\ell=j+1}^{n-1}\left(1-\frac{1-\alpha_\ell}{\ell}\right)+\prod_{j=1}^{n-1}\left(1-\frac{1-\alpha_j}{j}\right)\\ &= \sum_{j=1}^{n-1}\alpha_j\prod_{\ell=j+1}^{n-1}\left(1-\frac{1-\alpha_\ell}{\ell}\right)+\alpha_1\prod_{j=2}^{n-1}\left(1-\frac{1-\alpha_j}{j}\right). \qquad (9)\end{aligned}$$

Observe that $f(j,k) = 0$ if $j < k$. Therefore, for $k \geq 2$, (5) and (7) produce

$$\begin{aligned} f(n,k) &= \sum_{j=1}^{n-1} \frac{(1-\alpha_j)(k-1)f(j,k-1)}{j} \prod_{\ell=j+1}^{n-1} \left(1 - \frac{(1-\alpha_\ell)k}{\ell}\right) \\ &= \sum_{j=k-1}^{n} \frac{(1-\alpha_j)(k-1)f(j,k-1)}{j} \prod_{\ell=j+1}^{n-1} \left(1 - \frac{(1-\alpha_\ell)k}{\ell}\right) \\ &= \sum_{j=k}^{n} \frac{(1-\alpha_j)(k-1)f(j,k-1)}{j} \prod_{\ell=j+1}^{n-1} \left(1 - \frac{(1-\alpha_\ell)k}{\ell}\right) \\ &\quad + (1-\alpha_{k-1})f(k-1,k-1)\alpha_k \prod_{\ell=k+1}^{n-1} \left(1 - \frac{(1-\alpha_\ell)k}{\ell}\right). \end{aligned} \tag{10}$$

3.1. The case $r = 1$

Using (1) and Lemma 1 in (9) produces

$$\begin{aligned} f(n,1) &= \sum_{j=1}^{n-1} \left[\theta + O\left(j^{-\epsilon}\right)\right] \left(\frac{j}{n}\right)^{1-\theta} \left(1 + O\left(j^{-\epsilon'}\right)\right) + O\left(n^{\theta-1}\right) \\ &= \sum_{j=1}^{n-1} \theta \left(\frac{j}{n}\right)^{1-\theta} \left(1 + O\left(j^{-\epsilon'}\right)\right) + O\left(n^{\theta-1}\right) \\ &= \frac{\theta n}{2-\theta} + O\left(n^{1-\epsilon'}\right). \end{aligned}$$

We have shown that (2) holds when $k = 1$.

Suppose, now, that (2) holds for all $f(n,k-1)$. We will show that (2) holds for $f(n,k)$ holds as well. Using (1) and Lemma 1 in (10), we have

$$\begin{aligned} f(n,k) &= \sum_{j=k}^{n} \frac{\left[1-\theta-O(j^{-\epsilon})\right](k-1)\left[M_{k-1}j + O(j^{1-\epsilon'})\right]}{j} \left(\frac{j}{n}\right)^{k-k\theta} \left(1 + O\left(j^{-\epsilon'}\right)\right) \\ &\quad + O\left(n^{k\theta-k}\right) \\ &= (1-\theta)(k-1)M_{k-1}n^{k\theta-k} \sum_{j=k}^{n} j^{k-k\theta} \left(1 + O\left(j^{-\epsilon'}\right)\right) + O\left(n^{k\theta-k}\right) \\ &= \frac{(1-\theta)(k-1)}{k-k\theta+1} M_{k-1}n + O\left(n^{1-\epsilon'}\right) \\ &= M_k n + O\left(n^{1-\epsilon'}\right), \end{aligned}$$

proving the inductive step. Therefore, by induction, (2) holds for all $k \geq 1$.

3.2. The case $r \in (-1,1)$

Now, (1) and Lemma 1 in (9) give

$$\begin{aligned} f(n,1) &= \sum_{j=1}^{n-1} \left[\theta j^{r-1} + O\left(j^{r-1-\epsilon}\right)\right] \left(\frac{j}{n}\right) \left(1 + O\left(j^{-\epsilon'}\right)\right) + O\left(n^{-1}\right) \\ &= \theta n^{-1} \sum_{j=2}^{n} j^r \left(1 + O\left(j^{-\epsilon'}\right)\right) + O\left(n^{-1}\right) \\ &= \frac{\theta n^r}{r+1} + O\left(n^{r-\epsilon'}\right), \end{aligned}$$

establishing (3) for $k = 1$.

Define N_k for $k \geq 1$ by $N_k = \theta(k-1)!/(r+1)^{(k)}$. Now, suppose that $r \in (-1,1)$ and that (3) holds for $f(n, k-1)$. We have

$$\begin{aligned} f(n,k) &= \sum_{j=k}^{n} \frac{\left[1 - O(j^{r-1})\right](k-1)\left[N_{k-1} j^r + O\left(j^{r-\epsilon'}\right)\right]}{j} \left(\frac{j}{n}\right)^k \left(1 + O\left(j^{-\epsilon'}\right)\right) \\ &\quad + O\left(n^{-k}\right) \\ &= (k-1) N_{k-1} n^{-k} \sum_{j=k}^{n} j^{r+k-1} \left(1 + O\left(j^{-\epsilon'}\right)\right) + O\left(n^{-k}\right) \\ &= \frac{k-1}{r+k} N_{k-1} n^r + O\left(n^{r-\epsilon'}\right) \\ &= N_k n^r + O\left(n^{r-\epsilon'}\right), \end{aligned}$$

proving the inductive step and therefore showing that (3) is valid for all $k \geq 1$.

References

[1] Albert, R., Barabási, A.-L., 2002. Statistical mechanics of complex networks. Review of Modern Physics 74, 47–97.

[2] Aldous, D. J., 1985, Exchangeability and related topics. In P. Hennequin, editor, École d'été de probabilités de Saint-Flour, XIII-1983, pages 1–198, Springer, Berlin. Lecture Notes in Mathematics 1117.

[3] Arratia, R., Barbour .A D., Tavaré, S., 2003. Logarithmic combinatorial structures: A probabilistic approach. European Mathematical Society Publishing House.

[4] Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

[5] Chung, F., Lu, L., 2006. Concentration inequalities and martingale inequalities: A survey. Internet Mathematics 3, 79–127.

[6] Eriksson, K., Sjöstrand J., 2012. Limiting shapes of birth-and-death processes on Young diagrams. Adv. Appl. Math 48, 575-602.

[7] Erlihson, M. M., Granovsky, B. L., 2008. Limit shapes of Gibbs distributions on the set of integer partitions: The expansive case. Annales de l'Institut henri Poincaré - Probabilités et Statistiques 44, 915–945.

[8] Ewens, W. J., 2004. Mathematical Population Genetics, 2nd edition. Springer.

[9] Graham, R. L., Knuth, D. E., Patashnik, O., 1989. Concrete Mathematics. Addison-Wesley.

[10] Hoppe, F. M., 1984. Polya-like urns and the Ewens sampling formula. J. Math. Biol. 20, 91–99.

[11] Johnson, N., Kotz S., 1977. Urn Models and Their Applications: An Approach to Modern Discrete Probability Theory. Wiley, New York.

[12] Simon, H., 1955. On a class of skew distribution functions. Biometrika 42, 425–440.

[13] Watterson, G. A., 1974. Models for the logarithmic species abundance distributions. Theoretical Population Biology 6, 217–250.