Adv. Appl. Prob. **42,** 994–1012 (2010)
Printed in Northern Ireland

GAUSSIAN PHASES IN GENERALIZED COUPON COLLECTION

HOSAM M. MAHMOUD,* *The George Washington University*

Abstract

In this paper we consider a generalized coupon collection problem in which a customer repeatedly buys a random number of distinct coupons in order to gather a large number n of available coupons. We address the following question: How many different coupons are collected after $k = k_n$ draws, as $n \to \infty$? We identify three phases of k_n: the sublinear, the linear, and the superlinear. In the growing sublinear phase we see $o(n)$ different coupons, and, with true randomness in the number of purchases, under the appropriate centering and scaling, a Gaussian distribution is obtained across the entire phase. However, if the number of purchases is fixed, a degeneracy arises and normality holds only at the higher end of this phase. If the number of purchases have a fixed range, the small number of different coupons collected in the sublinear phase is upgraded to a number in need of centering and scaling to become normally distributed in the linear phase with a different normal distribution of the type that appears in the usual central limit theorems. The Gaussian results are obtained via martingale theory. We say a few words in passing about the high probability of collecting nearly all the coupons in the superlinear phase. It is our aim to present the results in a way that explores the critical transition at the 'seam line' between different Gaussian phases, and between these phases and other nonnormal phases.

Keywords: Urn model; random structure; martingale; central limit theorem; coupon collection

2010 Mathematics Subject Classification: Primary 60C05; 60F05; 05A05
Secondary 60G42

1. Classical coupon collection

Combinatorial problems underlying coupon collection procedures became popular in the 1930s, when the Dixie Cup ice cream company sold ice cream cups with a cardboard cover that had hidden on the underside a coupon (such as the picture of a well-known baseball player). The idea in this marketing strategy was to encourage fans, mostly young boys, to go for more purchases to complete a set of pictures and receive some kind of reward. The underlying structure is the following. A purchase of a certain product is awarded with one of n distinct equally likely coupons. When a coupon is collected, the purchaser keeps it, and the company replaces the product in the market. The classical problem deals with the waiting time (in terms of the number of purchases) until a purchaser collects all n different coupons.

Coupon collection can be visualized in terms of schemes of drawing balls from urns. For example, we can think of the problem as an urn containing n balls, of n different colors, sampled repeatedly *with replacement*. The classical question is: How many draws are necessary to observe the n colors? A second helpful scheme is that of an urn containing white and red balls:

Received 8 April 2009; revision received 24 September 2010.
* Postal address: Department of Statistics, The George Washington University, Washington, DC 20052, USA.
Email address: hosam@gwu.edu

at any stage, the white balls represent unobserved coupons and the red balls represent coupons that have already appeared in a previous draw. At each stage a ball is picked; if the sampled ball is red, we put it back in the urn. If the sampled ball is white, paint it red and put it back in the urn. In this scheme the classical question becomes: How many draws are necessary for an all white urn to become all red for the first time?

2. Generalized coupon collection and questions about urn composition at different stages

Coupon collection problems have a long history and can be traced back to Laplace and De Moivre. Stadje (1990) provided a good background to the history of the problem and posed additional nonclassic questions. We will focus on the generalized coupon collection problem and the more recent bibliography related to it.

A generalized form of the classical coupon collector's problem assumes that the customer purchases a random number, $S \geq 1$, of items each time and that the company guarantees that the S associated coupons are distinct. The customer obtains S coupons at each purchase, of which some or all may already be in his/her possession. The classical coupon collection problem corresponds to the case in which $S \equiv 1$. The generalized problem was addressed in Sellke (1995) and Adler and Ross (2001). For a review of its scope, see Kobza *et al.* (2007).

The average waiting time till all n coupons are collected has been investigated in the literature. Pólya (1930) tackled the problem and provided a formula, for fixed S, for the average waiting time till all the coupons are collected. For a large number, n, of coupons, the formula is unwieldy—it is difficult to compute as it contains a sum of very large terms with alternating signs, for which Pólya (1930) worked out an alternative approximation. Johnson and Sellke (2010) and Ivchenko (1998) dealt with a more general setup, in which at each purchase a random number, S, of items is obtained, and the realizations of S at each purchase are independent, identically distributed random variables. For this generalized version, Johnson and Sellke (2010) and Ivchenko (1998) also obtained exact formulae and various approximations for the average waiting time.

In our investigation we consider the case of *random* S, $1 \leq S \leq n$, but we address a set of issues other than waiting times. The application we have in mind is that of a family with three boys, say, who purchases three ice cream cups most of the time. Occasionally, one of the boys may want something else or the boys are in the company of a few friends and the mother treats everybody. A random S distributed on a small range suits this model. For transparency, we will present the results for S with a distribution on $\{1, 2, \ldots, s\}$ for fixed s, and only mention in the concluding remarks possible extensions that cover larger ranges.

We ask about the number of different coupons collected after a certain number of purchases. This question is of interest to market planners. For example, knowing that there are n coupons, a family may be willing to allocate a budget for at most $4n$ purchases, hoping that they will collect most of the coupons.

An underlying urn has the following scheme. White balls represent uncollected coupons and red balls represent collected coupons. Initially, there are n white balls in the urn. A sample of size S is taken out of the urn *without replacement* and the white balls in it are recolored red. The entire sample is then returned to the urn. The process is then repeated k times, and at each step an independent copy of S is generated for the sample size. We refer to the picking out of a sample of size S as a *sample draw*. In terms of the urn, we are interested in the number of red balls present in the urn after a certain number of sample draws. In other words, what is the urn composition after say k sample draws? We allow k to depend on n and the question we ask is: How many red balls are in the urn after $k := k_n$ sample draws, as $n \to \infty$?

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

We point out here that the set of S balls is drawn without replacement, meaning that the S balls are obtained randomly, one at a time, and an extracted ball is kept out of the urn until all the other members of a sample draw are taken from the urn. In other words, the sample is obtained by drawing a ball at random from among the n balls in the urn and setting it aside, then a second ball is drawn at random from among the remaining $n-1$ balls in the urn and set aside, and so forth until a sample of size S is obtained, at which point the white balls in the urn are colored red and the whole sample is put back in the urn.

We will identify three phases of k_n:

(a) the sublinear phase, when $k_n = o(n)$;

(b) the linear phase, when $k_n \sim \alpha_n n$ for some $\alpha_n > 0$ of a magnitude bounded from above and below;

(c) the superlinear phase, when $n = o(k_n)$.

Trivially, for the sublinear phase, the different coupons collected are relatively few. When S has genuine variability (positive variance), there is enough dispersion via the variability in the sample, when k_n goes sublinearly to ∞, to warrant normality under appropriate centering and scaling. However, when S is deterministic, variability for normality comes from an extended number of draws, and k_n has to be sufficiently high to achieve this. For fixed S, normality (under appropriate centering and scaling) is reached at the upper end of the sublinear phase, when $\sqrt{n} = o(k_n)$, and $k_n = o(n)$ still. In the linear phase centering and scaling by $\sqrt{n v_n}$ (where $n v_n$, with v_n nonzero but $O(1)$, is the asymptotic variance), as is usually the case in central limit theorems, give a different limiting Gaussian distribution. In the superlinear phase almost all the coupons are collected with high probability. In all the Gaussian phases identified the results are proved via martingale theory. We are able to extend several of these results to cases with a large (deterministic and random) number of purchases.

The rest of this paper has the following organization. Section 3 contains a brief description of the notation used throughout. In Section 4 we set up exact formulae, starting from an exact stochastic recurrence and ending with an exact calculation of the mean and variance of the number of white balls after n sample draws. In Section 5 we derive the underlying martingale. In Section 6 we discuss the three phases, the sublinear, the linear, and the superlinear, with a subsection devoted to each phase. The concluding remarks in Section 7 give interpretations for how the results in different phases conjoin at the 'seam lines'. The last of the remarks connect this work to areas of research in graph theory and occupancy problems.

3. Notation

At each sample draw a set of S balls is drawn from the urn, with $1 \le S \le s_n \le n$, and S has a discrete distribution on the set $\{1, 2, \ldots, s_n\}$. The random sample size S is independent of the urn content and all past sample sizes. In other words, we generate a sequence $S_1, S_2, \ldots$ of independent random variables having the distribution of S, and use S_i as the sample size in the ith stage.

We will give the full exposition for $s_n = s$ fixed, and only mention in the concluding remarks extensions to cases with increasing s_n. For fixed $s_n = s$, we denote the mean and variance of S by μ_S and σ_S^2.

Throughout, we will use the following standard probability notation. We denote the normally distributed random variate with mean 0 and variance ν^2 by $\mathcal{N}(0, \nu^2)$. We use the symbols '$\xrightarrow{\mathrm{D}}$', '$\xrightarrow{\mathrm{P}}$', and '$\xrightarrow{\mathrm{a.s.}}$' respectively for convergence in distribution, convergence in probability, and

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

almost-sure convergence, and use '$\stackrel{\mathcal{L}}{=}$' to denote exact equality in law. The notation $o_{\mathcal{L}_1}(g(n))$ will stand for a sequence of random variables that is $o(g(n))$ in the $\mathcal{L}_1$ norm, that is, when we describe a sequence of random variables X_n to be $o_{\mathcal{L}_1}(g(n))$, we mean that $\mathrm{E}[|X_n|]/g(n) \to 0$.

Let Hypergeo(N, m, a) be a hypergeometric random variable that represents the number of amber balls in a sample of m balls drawn at random (all subsets of size m being equally likely) from an urn containing a total of N amber and black balls, of which a are amber. The mean and variance for this standard distribution are given by

$$\mathrm{E}[\mathrm{Hypergeo}(N, m, a)] = \frac{am}{N}, \tag{1}$$

$$\mathrm{var}[\mathrm{Hypergeo}(N, m, a)] = \frac{am(N-a)(N-m)}{N^2(N-1)}. \tag{2}$$

Unless otherwise stated, all asymptotics will mean asymptotic equivalents and bounds as $n \to \infty$. The number $n/(n - \mu_S)$ will appear often, and we will give it the designation ρ_n. We will repeatedly use well-known facts about ρ_n^{yn} for $y > 0$, such as the fact that ρ_n^{yn} is asymptotically $\mathrm{e}^{\mu_S y} + O(1/n)$.

We will also need the backward difference operator ∇, which, when applied to a function $h(i)$, with integer argument i, gives the difference between two successive steps, that is, $\nabla h(i) = h(i) - h(i-1)$. The indicator $\mathbf{1}_{\mathcal{E}}$ is a function of a sample space that assumes the value 1 if $\mathcal{E}$ occurs and 0 otherwise.

4. Exact moments

In the generalized coupon collection problem there are initially n balls in the urn. We sample S, $1 \le S \le s$, balls at a time and return them to the urn with all white balls in the sample recolored red. Let R_j be the number of red balls (collected coupons) and let W_j be the number of white balls (uncollected coupons) after j such sample draws. For any $j \ge 0$, we have $R_j + W_j = n$. There is stochastic dependence between W_{j-1} and W_j. After j sample draws, the number of white balls in the urn is equal to the number of white balls after the $(j-1)$th draw minus ω_j, the number of white balls that are recolored red. Given S and W_{j-1} (which are independent), the number of white balls appearing in the jth sample is distributed as Hypergeo(n, S, W_{j-1}). Hence,

$$W_j = W_{j-1} - \omega_j, \tag{3}$$

with

$$(\omega_j \mid W_{j-1}, S) \stackrel{\mathcal{L}}{=} \mathrm{Hypergeo}(n, S, W_{j-1}). \tag{4}$$

It follows from the stochastic recurrence (3) and the conditional hypergeometric distribution (4) of ω_j (the mean of which is given in (1)) that

$$\begin{aligned}
\mathrm{E}[W_j] &= \mathrm{E}[W_{j-1}] - \mathrm{E}[\omega_j] \\
&= \mathrm{E}[W_{j-1}] - \mathrm{E}[\mathrm{E}[\omega_j \mid W_{j-1}, S]] \\
&= \mathrm{E}[W_{j-1}] - \mathrm{E}\left[\frac{SW_{j-1}}{n}\right] \\
&= \mathrm{E}[W_{j-1}] - \frac{1}{n}\mathrm{E}[S]\,\mathrm{E}[W_{j-1}] \\
&= \left(1 - \frac{\mu_S}{n}\right)\mathrm{E}[W_{j-1}]
\end{aligned}$$

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

$$= \left(1 - \frac{\mu_S}{n}\right)^2 \mathrm{E}[W_{j-2}]$$
$$= \cdots$$
$$= \left(1 - \frac{\mu_S}{n}\right)^j n, \tag{5}$$

where in the final step we used the initial condition $W_0 = n$.

The second moment, and subsequently the variance, also follows from (3) in its squared form:

$$W_j^2 = W_{j-1}^2 - 2\omega_j W_{j-1} + \omega_j^2.$$

Upon taking the conditional expectation (given W_{j-1} and S), we obtain

$$\mathrm{E}[W_j^2 \mid W_{j-1}, S] = W_{j-1}^2 - 2W_{j-1}\,\mathrm{E}[\omega_j \mid W_{j-1}, S] + \mathrm{E}[\omega_j^2 \mid W_{j-1}, S].$$

Then, $(\omega_j \mid W_{j-1}, S)$ has the distribution of a Hypergeo(n, S, W_{j-1}) random variable, for which the mean and variance are given by the standard forms (1) and (2). If we substitute these forms into the last equality and simplify, we obtain

$$\mathrm{E}[W_j^2 \mid W_{j-1}, S] = \frac{(n-S)(n-S-1)}{n(n-1)} W_{j-1}^2 + \frac{S(n-S)}{n(n-1)} W_{j-1}.$$

The second unconditional moment follows from the last equation by taking its expectation, yielding the recurrence

$$\mathrm{E}[W_j^2] = \mathrm{E}\left[\frac{(n-S)(n-S-1)}{n(n-1)}\right]\mathrm{E}[W_{j-1}^2] + \mathrm{E}\left[\frac{S(n-S)}{n(n-1)}\right]\mathrm{E}[W_{j-1}],$$

which has the solution

$$\mathrm{E}[W_j^2] = n\left[(n-1)\left(\frac{(n-\mu_S)(n-\mu_S-1)+\sigma_S^2}{n(n-1)}\right)^j + \left(\frac{n-\mu_S}{n}\right)^j\right].$$

Therefore, the variance is

$$\mathrm{var}[W_j] = n\left[(n-1)\left(\frac{(n-\mu_S)(n-\mu_S-1)+\sigma_S^2}{n(n-1)}\right)^j + \left(\frac{n-\mu_S}{n}\right)^j\right] - \left(\frac{n-\mu_S}{n}\right)^{2j} n^2. \tag{6}$$

5. A martingale underlying the urn scheme

Let $\mathcal{F}_j$ be the sigma field generated by the first j sample draws. This sigma field contains all the information that can be gleaned from j sample draws. With $(\omega_j \mid W_{j-1}, S)$ having the distribution of Hypergeo(n, S, W_{j-1}), with average SW_{j-1}/n (as given in (1)), we obtain

$$\mathrm{E}[W_j \mid \mathcal{F}_{j-1}] = W_{j-1} - \mathrm{E}[\omega_j \mid \mathcal{F}_{j-1}] = \left(1 - \frac{\mu_S}{n}\right) W_{j-1}.$$

It then immediately follows that

$$Y_j = \left(\frac{n}{n-\mu_S}\right)^j W_j = \rho_n^j W_j$$

is a martingale.

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

The fact that Y_j is a martingale is key to proving central limit theorems in all the Gaussian phases. We will deal with the centered martingale

$$\tilde{Y}_j = Y_j - n$$

(which has mean 0) to employ the martingale central limit theorem, which requires calculations on a zero-mean martingale. Sufficient conditions for the central limit theorem for a zero-mean martingale $X_{j,n}$ are the conditional Lindeberg condition and the conditional variance condition on the martingale differences $\nabla X_{j,k_n} = X_{j,k_n} - X_{j-1,k_n}$; see Theorem 3.2 and Corollary 3.1 of Hall and Hyde (1980, p. 58).

Specifically, in our case, the conditional Lindeberg condition requires that, for some positive increasing sequence λ_n and all $\varepsilon > 0$,

$$U_n := \sum_{j=1}^{k_n} \mathrm{E}\left[\left(\frac{\nabla\tilde{Y}_j}{\lambda_n}\right)^2 \mathbf{1}_{\{|\nabla\tilde{Y}_j/\lambda_n|>\varepsilon\}} \;\middle|\; \mathcal{F}_{j-1}\right] \xrightarrow{\mathrm{P}} 0, \tag{7}$$

and a Z-conditional variance condition requires that

$$V_n := \sum_{j=1}^{k_n} \mathrm{E}\left[\left(\frac{\nabla\tilde{Y}_j}{\lambda_n}\right)^2 \;\middle|\; \mathcal{F}_{j-1}\right] \xrightarrow{\mathrm{P}} Z. \tag{8}$$

When both conditions hold, the sum $\sum_{j=1}^{k_n} \nabla\tilde{Y}_j/\lambda_n = (Y_{k_n} - Y_0)/\lambda_n = (Y_{k_n} - n)/\lambda_n$ converges to a mixture of normally distributed random variables with characteristic function $\mathrm{E}[\exp(-Zt^2/2)]$. When Z is the constant c^2, the mixture is simply the $\mathcal{N}(0, c^2)$ random variable.

To derive a martingale central limit theorem in any of the phases, we need to identify the appropriate scale λ_n for that phase. For calculations involved in Lindeberg's conditional condition, we need $\mathrm{E}[(\nabla\tilde{Y}_j)^2 \mid \mathcal{F}_{j-1}]$ (see the definition of V_n in (8)); we find that

$$\begin{aligned}
\mathrm{E}[(\nabla\tilde{Y}_j)^2 \mid \mathcal{F}_{j-1}] &= \mathrm{E}[(\rho_n^j W_j - \rho_n^{j-1} W_{j-1})^2 \mid \mathcal{F}_{j-1}] \\
&= \left(\mathrm{E}\left[\frac{(n-S)(n-S-1)}{n(n-1)}\right]\rho_n^{2j} - 2\,\mathrm{E}\left[\frac{n-S}{n}\right]\rho_n^{2j-1} + \rho_n^{2j-2}\right) W_{j-1}^2 \\
&\quad + \mathrm{E}\left[\frac{S(n-S)}{n(n-1)}\right]\rho_n^{2j} W_{j-1} \\
&= \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{n^2(n-1)}\rho_n^{2j} W_{j-1}^2 + \frac{\mu_S n - \mu_S^2 - \sigma_S^2}{n(n-1)}\rho_n^{2j} W_{j-1}.
\end{aligned}$$

Summarizing, we construct V_n as

$$V_n = \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{\lambda_n^2 n^2(n-1)} \sum_{j=1}^{k_n} \rho_n^{2j} W_{j-1}^2 + \frac{\mu_S n - \mu_S^2 - \sigma_S^2}{\lambda_n^2 n(n-1)} \sum_{j=1}^{k_n} \rho_n^{2j} W_{j-1}. \tag{9}$$

6. Phases during long-term drawing

Imagine indefinitely drawing from the urn according to the rules. Many stochastic paths will deplete the white balls after some time, and the urn will remain all red after a number of sample draws. We will see that, as the drawing continues, the process experiences different phases.

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

6.1. The sublinear phase

Suppose that $0 \le k_n = o(n)$. Trivially, at most $sk_n = o(n)$ white balls can turn red, and $n - sk_n \le W_{k_n} \le n$. Thus,

$$W_j = n + O(k_n) \tag{10}$$

for each $0 \le j \le k_n$, and

$$\frac{W_{k_n}}{n} \overset{\text{a.s.}}{\to} 1. \tag{11}$$

Lemma 1. *In the sublinear phase the absolute differences $|\nabla \tilde{Y}_j|$ are uniformly bounded for all n greater than some integer N_0.*

Proof. Consider $1 \le j \le k_n = o(n)$. For large enough n (greater than some $N_0 > 2s$),

$$\rho_n^{j-1} = \left(\frac{n}{n-\mu_S}\right)^{j-1} \le 2.$$

Take $n > N_0$, and write the absolute differences as

$$\begin{aligned} |\nabla \tilde{Y}_j| &= |(Y_j - n) - (Y_{j-1} - n)| \\ &= \rho_n^{j-1} |\rho_n (W_{j-1} - \omega_j) - W_{j-1}| \\ &\le 2|(\rho_n - 1) W_{j-1} - \rho_n \omega_j| \\ &\le 2\left(\left(\frac{n}{n-\mu_S} - 1\right) W_{j-1} + \frac{n}{n-\mu_S}\omega_j\right). \end{aligned}$$

The number of white balls at any stage is at most n, and the change (the reduction by ω_j) is at most s. Then it follows that

$$|\nabla \tilde{Y}_j| \le 2\frac{2sn}{n-\mu_S} \le 8s.$$

This completes the proof.

By appropriate centering and scaling, we can refine the strong law in (11) and find that, for $\sigma_S^2 > 0$, its rate of convergence is a Gaussian random variable across the entire growing sublinear phase ($k_n \to \infty$ and $k_n = o(n)$). For the degenerate case of fixed purchases each time ($S = s$, that is, $\sigma_S^2 = 0$), the argument breaks down for k_n of the order $\sqrt{n}$ (or lower), and a different Gaussian random variable takes over as the limit when

$$\sqrt{n} = o(k_n) \quad \text{and} \quad k_n = o(n);$$

we call this phase the *upper sublinear phase*. By contrast we call the rest of the sublinear range the *lower sublinear phase*. In the lower sublinear phase k_n may grow to ∞, or stay bounded. We call the lower sublinear phase in which $k_n \to \infty$ the *growing lower sublinear phase*. We refer to the growing lower sublinear phase and the upper sublinear phase as the *growing sublinear phase*.

Theorem 1. *Let R_{k_n} be the number of collected coupons (red balls in the urn) after k_n purchases (sample draws from the urn), where k_n is in the growing sublinear phase. Then,*

(a) *if $\sigma_S^2 > 0$,*

$$\frac{R_{k_n} - n + (1 - \mu_S/n)^{k_n} n}{\sqrt{k_n}} \overset{\text{D}}{\to} \mathcal{N}(0, \sigma_S^2);$$

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

(b) *in the case where the number of purchases at each step is fixed at s (the sample draws are of fixed size s, with $\sigma_S^2 = 0$), a Gaussian law holds in the upper sublinear phase:*

$$\frac{R_{k_n} - n + (1 - s/n)^{k_n} n}{k_n/\sqrt{n}} \xrightarrow{\text{D}} \mathcal{N}\left(0, \frac{1}{2}s^2\right).$$

Mikhaĭlov (1980) considered similar cases to Theorem1(b) using the method of moments. We present a proof via martingales, which can be generalized to the case of a large number of purchases (deterministic but growing with n).

Proof of Theorem 1. (a) Assume that $\sigma_S^2 > 0$, and that k_n grows to ∞ in any sublinear manner ($k_n = o(n)$). For this sublinear phase, take the scale factor $\lambda_n = \sqrt{k_n}$. Recall the expressions for U_n (cf. (7)) and V_n (cf. (8)). The proof will be complete if we show that U_n converges to 0 in probability and V_n converges to σ_S^2 in probability.

For the conditional Lindeberg condition, we have the uniform upper bound of $8s$ for $|\nabla \tilde{Y}_j|$ for all n greater than some $N_0 > 2s$ (see Lemma 1). Therefore, for any $\varepsilon > 0$,

$$U_n = \sum_{j=1}^{k_n} \mathrm{E}\left[\left(\frac{\nabla \tilde{Y}_j}{\sqrt{k_n}}\right)^2 \mathbf{1}_{\{|\nabla \tilde{Y}_j/\sqrt{k_n}| > \varepsilon\}} \;\middle|\; \mathcal{F}_{j-1}\right],$$

where the sets $\{|\nabla \tilde{Y}_j| > \varepsilon\sqrt{k_n}\}$ are all empty, for all n greater than some $n_0(\varepsilon) > N_0$. For large n, we have

$$\begin{aligned} U_n &= \sum_{j=1}^{n_0(\varepsilon)} \mathrm{E}\left[\left(\frac{\nabla \tilde{Y}_j}{\sqrt{k_n}}\right)^2 \mathbf{1}_{\{|\nabla \tilde{Y}_j/\sqrt{k_n}| > \varepsilon\}} \;\middle|\; \mathcal{F}_{j-1}\right] \\ &\le \frac{1}{k_n} \sum_{j=1}^{n_0(\varepsilon)} \mathrm{E}[(\nabla \tilde{Y}_j)^2 \mid \mathcal{F}_{j-1}] \\ &\le \frac{64 s^2 n_0(\varepsilon)}{k_n} \\ &\to 0 \quad \text{as } n \to \infty. \end{aligned}$$

Hence, the conditional Lindeberg condition is verified in the entire growing sublinear phase.

In (9) replace W_{j-1} by the asymptotic equivalent in (10) to obtain

$$\begin{aligned} V_n &= \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{\lambda_n^2 n^2 (n-1)} \sum_{j=1}^{k_n} \rho_n^{2j} (n + O(k_n))^2 + \frac{\mu_S n - \mu_S^2 - \sigma_S^2}{\lambda_n^2 n (n-1)} \sum_{j=1}^{k_n} \rho_n^{2j} (n + O(k_n)) \\ &= \frac{\sigma_S^2}{k_n}\left(1 + O\left(\frac{k_n}{n}\right)\right) \sum_{j=1}^{k_n} \rho_n^{2j}. \end{aligned}$$

The geometric series $A_n = \sum_{j=1}^{k_n} \rho_n^{2j}$ can be asymptotically summed as follows:

$$\begin{aligned} A_n &= \frac{(n/(n-\mu_S))^{2k_n+2} - 1}{(n/(n-\mu_S))^2 - 1} - 1 \\ &= \frac{(n-\mu_S)^2}{\mu_S(2n-\mu_S)} (\mathrm{e}^{(2k_n+2)\ln(n/(n-\mu_S))} - 1) - 1 \end{aligned}$$

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

$$= \frac{(n-\mu_S)^2}{\mu_S(2n-\mu_S)}\left(\left(1+\frac{2\mu_S k_n}{n}+O\left(\frac{k_n^2}{n^2}\right)\right)-1\right)-1$$
$$= k_n + o(k_n). \tag{12}$$

It follows that

$$V_n = \frac{\sigma_S^2}{k_n}\left(1+O\left(\frac{k_n}{n}\right)\right)(k_n+o(k_n)) \to \sigma_S^2.$$

Hence, the σ_S^2-conditional variance condition is verified in the entire growing sublinear phase.

With both conditions checked, the martingale central limit theorem gives

$$\sum_{j=1}^{k_n}\left(\frac{\nabla\tilde{Y}_j}{\sqrt{k_n}}\right) = \frac{Y_{k_n}-Y_0}{\sqrt{k_n}} \xrightarrow{\text{D}} \mathcal{N}(0,\sigma_S^2).$$

Subsequently, we write

$$\frac{\rho_n^{k_n}W_{k_n}-\rho_n^0 n}{\sqrt{k_n}} = \frac{(n/(n-\mu_S))^{k_n}W_{k_n}-n}{\sqrt{k_n}} \xrightarrow{\text{D}} \mathcal{N}(0,\sigma_S^2).$$

Using the fact that $(n/(n-\mu_S))^{-k_n}$ converges to 1 in the growing sublinear phase and an application of Slutsky's multiplicative theorem (see Karr (1993, p. 147)), we obtain

$$\frac{W_{k_n}-(n/(n-\mu_S))^{-k_n}n}{\sqrt{k_n}} \xrightarrow{\text{D}} \mathcal{N}(0,\sigma_S^2).$$

Theorem 1(a) follows in its stated form from the relation $R_{k_n}+W_{k_n}=n$.

(b) Suppose that $S=s$ deterministically (that is, $\sigma_S^2=0$). Assume that $\sqrt{n}=o(k_n)$ and $k_n=o(n)$. Recall the expressions for U_n (cf. (7)) and V_n (cf. (8)). For this sublinear phase, take the scale factor $\lambda_n=k_n/\sqrt{n}$. The proof will be complete if we show that U_n converges to 0 in probability and V_n converges to $s^2/2$ in probability.

For the conditional Lindeberg condition, we have the uniform upper bound of $8s$ for $|\nabla\tilde{Y}_j|$ for all n greater than some $N_0>2s$ (see Lemma 1). Therefore, for any $\varepsilon>0$,

$$U_n = \sum_{j=1}^{k_n}\mathrm{E}\left[\left(\frac{\nabla\tilde{Y}_j}{k_n/\sqrt{n}}\right)^2\mathbf{1}_{\{|\nabla\tilde{Y}_j/(k_n/\sqrt{n})|>\varepsilon\}}\;\middle|\;\mathcal{F}_{j-1}\right],$$

where the sets $\{|\nabla\tilde{Y}_j|>\varepsilon k_n/\sqrt{n}\}$ are all empty, for all n greater than some $n_0'(\varepsilon)>N_0$. For large n, we have

$$\begin{aligned}U_n &= \sum_{j=1}^{n_0'(\varepsilon)}\mathrm{E}\left[\left(\frac{\nabla\tilde{Y}_j}{k_n/\sqrt{n}}\right)^2\mathbf{1}_{\{|\nabla\tilde{Y}_j/(k_n/\sqrt{n})|>\varepsilon\}}\;\middle|\;\mathcal{F}_{j-1}\right]\\ &\le \frac{n}{k_n^2}\sum_{j=1}^{n_0'(\varepsilon)}\mathrm{E}[(\nabla\tilde{Y}_j)^2\mid\mathcal{F}_{j-1}]\\ &\le \frac{64s^2n_0'(\varepsilon)n}{k_n^2}\\ &\to 0 \quad \text{as } n\to\infty.\end{aligned}$$

Hence, the conditional Lindeberg condition is verified in the upper sublinear phase.

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

An asymptotic analysis of the exact variance formula (6) shows that in this upper sublinear phase the variance is of order k_n^2/n. By Chebyshev's inequality, for any $0 \le j \le k_n$ and any $\varepsilon > 0$, we have

$$\begin{aligned}\mathrm{P}(|W_j - \mathrm{E}[W_j]| > \varepsilon k_n) &\le \frac{\mathrm{var}[W_j]}{\varepsilon^2 k_n^2} \\ &= \frac{O(k_n^2/n)}{\varepsilon^2 k_n^2} \\ &= O\left(\frac{1}{n}\right) \\ &\to 0.\end{aligned}$$

Whence, $(W_j - \mathrm{E}[W_j])/k_n \xrightarrow{\mathrm{P}} 0$, and we have the asymptotic representation

$$W_j = \left(1 - \frac{s}{n}\right)^j n + o_{\mathrm{P}}(k_n) = n - sj + o_{\mathrm{P}}(k_n) \tag{13}$$

for all $0 \le j \le k_n$.

In (9) replace W_{j-1} by the asymptotic equivalent in (13) to obtain

$$\begin{aligned}V_n &= \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{\lambda_n^2 n^2 (n-1)} \sum_{j=1}^{k_n} \rho_n^{2j} (n - s(j-1) + o_{\mathrm{P}}(k_n))^2 \\ &\quad + \frac{\mu_S n - \mu_S^2 - \sigma_S^2}{\lambda_n^2 n (n-1)} \sum_{j=1}^{k_n} \rho_n^{2j} (n - s(j-1) + o_{\mathrm{P}}(k_n)) \\ &= \frac{s(n-s)}{(n-1)k_n^2} \left(s \sum_{j=1}^{k_n} j \rho_n^{2j} + o_{\mathrm{P}}(k_n) \sum_{j=1}^{k_n} \rho_n^{2j} \right);\end{aligned}$$

here there are two sums of geometric series type:

$$B_n = \sum_{j=1}^{k_n} j \rho_n^{2j} \quad \text{and} \quad A_n = \sum_{j=1}^{k_n} \rho_n^{2j},$$

with A_n already handled in (12), where it was shown that $A_n = O(k_n)$.

The factor B_n is more delicate to analyze owing to multiple cancellations that necessitate we go further with local expansions:

$$\begin{aligned}B_n &= \rho_n^2 \frac{\rho_n^{2k_n}(k_n(\rho_n^2 - 1) - 1) + 1}{(\rho_n^2 - 1)^2} \\ &= \frac{n^2}{s^2(2n-s)^2} \left(\left(\frac{n}{n-s} \right)^{2k_n} [sk_n(2n-s) - (n-s)^2] + (n-s)^2 \right) \\ &= \frac{n^2}{s^2(2n-s)^2} \left(\left(1 + \frac{2sk_n}{n} + \frac{s^2 k_n}{n^2} + \frac{2s^2 k_n^2}{n^2} + O\left(\frac{k_n}{n^3}\right) \right) \right. \\ &\qquad \left. \times [sk_n(2n-s) - (n-s)^2] + (n-s)^2 \right) \\ &= \tfrac{1}{2} k_n^2 + O(k_n).\end{aligned}$$

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

Putting it all together we see that

$$\begin{aligned} V_n &= \frac{s(n-s)}{(n-1)k_n^2}(sB_n + A_n o_{\mathrm{P}}(k_n)) \\ &= \frac{s}{k_n^2}(1+o(1))\left[s\left(\frac{1}{2}k_n^2 + O(k_n)\right) + o_{\mathrm{P}}(k_n^2)\right] \\ &\to \tfrac{1}{2}s^2. \end{aligned}$$

Hence, the $\frac{1}{2}s^2$-conditional variance condition is verified in the upper sublinear phase.

With both conditions checked, the martingale central limit theorem gives

$$\sum_{j=1}^{k_n}\left(\frac{\nabla\tilde{Y}_j}{k_n/\sqrt{n}}\right) = \frac{Y_{k_n} - Y_0}{k_n/\sqrt{n}} \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, \frac{1}{2}s^2\right).$$

We complete the proof of Theorem 1(b) with a few adjustments by Slutsky's theorem, similarly to those given at the end of the proof of part (a).

6.2. The linear phase

In the linear phase $k_n \sim \alpha_n n$ for some $\alpha_n > 0$ of a magnitude uniformly bounded from above and below, that is, for two positive constants, M_1 and M_2, and all n, $M_1 \le \alpha_n \le M_2$. The result in this linear phase is similar to that in the sublinear phase. We address this similarity in a few brief remarks in Section 7.

At this phase of the drawing, we have the asymptotic equivalents (as $n \to \infty$), following from (5) and (6),

$$\mathrm{E}[W_{k_n}] = \mathrm{e}^{-\mu_S \alpha_n} n + o(n) \tag{14}$$

and

$$\mathrm{var}[W_{k_n}] \sim n v_n + o(n), \tag{15}$$

where

$$v_n = \frac{\mathrm{e}^{\mu_S \alpha_n} + \alpha_n(\sigma_S^2 - \mu_S) - 1}{\mathrm{e}^{2\mu_S \alpha_n}} = O(1).$$

We start with a first-order result for W_{k_n}.

Theorem 2. *For $k_n = \alpha_n n + o(n)$ for some $\alpha_n > 0$ of a magnitude bounded from above and below,*

$$\frac{W_{k_n}}{n\mathrm{e}^{-\mu_S \alpha_n}} \xrightarrow{\mathrm{P}} 1.$$

Proof. By Chebyshev's inequality,

$$\begin{aligned} \mathrm{P}(|W_{k_n} - \mathrm{E}[W_{k_n}]| \ge \varepsilon\, \mathrm{E}[W_{k_n}]) &\le \frac{\mathrm{var}[W_{k_n}]}{\varepsilon^2(\mathrm{E}[W_{k_n}])^2} \\ &\sim \frac{n v_n}{\varepsilon^2 \mathrm{e}^{-2\mu_S \alpha_n} n^2} \\ &\to 0 \quad \text{as } n \to \infty. \end{aligned}$$

Hence,

$$\frac{W_{k_n}}{\mathrm{E}[W_{k_n}]} \xrightarrow{\mathrm{P}} 1.$$

From the convergence $\mathrm{E}[W_{k_n}]/(n\mathrm{e}^{-\mu_S \alpha_n}) \to 1$, and Slutsky's theorem in its multiplicative form (cf. Karr (1993, p. 147)), the result follows.

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

Before we dwell on the proof of a central limit theorem for the number of coupons collected by the end of some linear phase, we need a technical lemma, which shows that W_{k_n} grows linearly with n like its mean. The purpose of this calculation is for later summation to verify Lindeberg's conditional condition.

Lemma 2. *Let W_{k_n} be the number of white balls in the urn after k_n draws, where $k_n = \alpha_n n + o(n)$ for some α_n such that $0 < M_1 \le \alpha_n \le M_2 < \infty$. Then*

$$W_{k_n} = \mathrm{e}^{-\mu_S \alpha_n} n + o_{\mathcal{L}_1}(n).$$

Proof. From the asymptotics of the mean and variance, as given in (14) and (15), for large n, we have

$$\begin{aligned}\mathrm{E}[(W_{k_n} - \mathrm{e}^{-\mu_S\alpha_n} n)^2] &= \mathrm{var}[W_{k_n}] + (\mathrm{E}[W_{k_n}] - \mathrm{e}^{-\mu_S\alpha_n} n)^2 \\ &= (\mathrm{e}^{\mu_S\alpha_n} + \alpha_n(\sigma_S^2 - \mu_S) - 1)\mathrm{e}^{-2\mu_S\alpha_n} n + o(n^2) \\ &= o(n^2).\end{aligned}$$

So, by the Cauchy–Schwarz inequality,

$$\mathrm{E}[|W_{k_n} - \mathrm{e}^{-\mu_S\alpha_n} n|] \le \sqrt{\mathrm{E}[(W_{k_n} - \mathrm{e}^{-\mu_S\alpha_n} n)^2]} = o(n),$$

which implies that

$$W_{k_n} = \mathrm{e}^{-\mu_S\alpha_n} n + o_{\mathcal{L}_1}(n).$$

This completes the proof.

Lemma 3. *For $j \sim yn \le M_2 n$ in the linear phase of drawing, the absolute differences $|\nabla \tilde{Y}_j|$ are uniformly bounded (in $n > N_0'$ for some integer $N_0' > 2s$).*

Proof. Suppose that $j \sim yn$, with $0 < y < M_2$, and write the absolute differences as

$$\begin{aligned}|\nabla \tilde{Y}_j| &= |(Y_j - n) - (Y_{j-1} - n)| \\ &= \rho_n^{j-1} |\rho_n (W_{j-1} - \omega_j) - W_{j-1}| \\ &\le 2\mathrm{e}^{y\mu_S} |(\rho_n - 1) W_{j-1} - \rho_n \omega_j| \quad (\text{for } n \text{ greater than some } N_0') \\ &\le 2\mathrm{e}^{M_2\mu_S} \left(\left(\frac{n}{n - \mu_S} - 1 \right) W_{j-1} + \frac{n}{n - \mu_S} \omega_j \right).\end{aligned}$$

The number of white balls at any stage is at most n, and the change (the reduction by ω_j) is at most s. Then it follows that

$$|\nabla \tilde{Y}_j| \le 2\mathrm{e}^{\mu_S M_2} \left(\frac{2sn}{n - \mu_S} \right) \le 8s\mathrm{e}^{\mu_S M_2}$$

for all $n > N_0'$.

Theorem 3. *Let R_{k_n} be the number of coupons collected (red balls in the urn) after k_n purchases (sample draws), where $k_n \sim \alpha_n n$ for some α_n such that $0 < M_1 \le \alpha_n \le M_2 < \infty$. Then,*

$$\frac{R_{k_n} - (1 - \mathrm{e}^{-\mu_S\alpha_n})n}{\sqrt{n((\mathrm{e}^{\mu_S\alpha_n} + \alpha_n(\sigma_S^2 - \mu_S) - 1)/\mathrm{e}^{2\mu_S\alpha_n})}} \xrightarrow{\mathrm{D}} \mathcal{N}(0, 1).$$

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

Proof. In this phase we take the scale factor λ_n to be $\sqrt{n v_n \mathrm{e}^{2\mu_S \alpha_n}}$, where

$$v_n = \frac{\mathrm{e}^{\mu_S \alpha_n} + \alpha_n(\sigma_S^2 - \mu_S) - 1}{\mathrm{e}^{2\mu_S \alpha_n}} = O(1).$$

Recall the expressions for U_n (cf. (7)) and V_n (cf. (8)). The proof will be complete if we show that U_n converges to 0 in probability and V_n converges to 1 in probability.

The conditional Lindeberg condition can be argued in view of the uniform bound of $8s\mathrm{e}^{\mu_S M_2}$ on the absolute differences $|\nabla \tilde{Y}_j|$ in the linear phase for $n > N_0'$; see Lemma 3. The set $\{|\nabla \tilde{Y}_j| > \varepsilon\sqrt{(\mathrm{e}^{\mu_S M_2} + M_2(\sigma_S^2 + \mu_S) + 1)n}\}$ is empty for all n greater than some $n_0''(\varepsilon) > N_0' > 2s$. The set $\{|\nabla \tilde{Y}_j| > \varepsilon\sqrt{n v_n \mathrm{e}^{2\mu_S \alpha_n}}\}$ is only a subset of it, so it is also empty for all n greater than some $n_0''(\varepsilon) > N_0' > 2s$. For large n, we have

$$\begin{aligned}
U_n &= \sum_{j=1}^{n_0''(\varepsilon)} \mathrm{E}\left[\left(\frac{\nabla \tilde{Y}_j}{\sqrt{n v_n \mathrm{e}^{2\mu_S \alpha_n}}}\right)^2 \mathbf{1}_{\{|\nabla \tilde{Y}_j/\sqrt{n v_n \mathrm{e}^{2\mu_S \alpha_n}}| > \varepsilon\}} \;\middle|\; \mathcal{F}_{j-1}\right] \\
&\le \frac{1}{n v_n \mathrm{e}^{2\mu_S \alpha_n}} \sum_{j=1}^{n_0''(\varepsilon)} \mathrm{E}[(\nabla \tilde{Y}_j)^2 \mid \mathcal{F}_{j-1}] \\
&\le \frac{64 s^2 \mathrm{e}^{2\mu_S M_2} n_0''(\varepsilon)}{n v_n \mathrm{e}^{2\mu_S \alpha_n}} \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

Hence, the conditional Lindeberg condition is verified.

The asymptotic equivalents in Lemma 2 apply only in the linear phase. However, before the linear phase the obvious bound n on W_{j-1} is sufficient for our purpose. More precisely, to asymptotically handle the sums in the conditional Lindeberg condition (going over the range of indexes 1 to $k_n \sim \alpha_n n$), let us break them up at some point near the beginning of the linear phase. Choose a small positive $\epsilon < M_1$ and break up the sums in V_n into sums going from 1 to $\lfloor \varepsilon n \rfloor - 1$ and sums starting at $\lfloor \varepsilon n \rfloor$ and ending at k_n. Applying the asymptotics of Lemma 2, we write (9) in the form

$$\begin{aligned}
V_n &= \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{\lambda_n^2 n^2 (n-1)} \sum_{j=1}^{\lfloor \varepsilon n \rfloor - 1} \rho_n^{2j} W_{j-1}^2 + \frac{\mu_S n - \mu_S^2 - \sigma_S^2}{\lambda_n^2 n (n-1)} \sum_{j=1}^{\lfloor \varepsilon n \rfloor - 1} \rho_n^{2j} W_{j-1} \\
&\quad + \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{\lambda_n^2 n^2 (n-1)} \sum_{j=\lfloor \varepsilon n \rfloor}^{k_n} (\rho_n^{2j} (\mathrm{e}^{-2\mu_S (j-1)/n} n^2 + o_{\mathcal{L}_1}(n^2))) \\
&\quad + \frac{\mu_S n - \mu_S^2 - \sigma_S^2}{\lambda_n^2 n (n-1)} \sum_{j=\lfloor \varepsilon n \rfloor}^{k_n} (\rho_n^{2j} (\mathrm{e}^{-\mu_S (j-1)/n} n + o_{\mathcal{L}_1}(n))) \\
&=: C_n + C_n' + D_n + H_n,
\end{aligned}$$

where

$$C_n = \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{n^3 (n-1) v_n \mathrm{e}^{2\mu_S \alpha_n}} \sum_{j=1}^{\lfloor \varepsilon n \rfloor - 1} \rho_n^{2j} W_{j-1}^2,$$

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

$$C_n' = \frac{\mu_S n - \mu_S^2 - \sigma_S^2}{n^2(n-1)v_n \mathrm{e}^{2\mu_S \alpha_n}} \sum_{j=1}^{\lfloor \varepsilon n \rfloor - 1} \rho_n^{2j} W_{j-1},$$

$$D_n = \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{n^3(n-1)v_n \mathrm{e}^{2\mu_S \alpha_n}} \sum_{j=\lfloor \varepsilon n \rfloor}^{k_n} (\rho_n^{2j}(\mathrm{e}^{-2\mu_S(j-1)/n} n^2 + o_{\mathcal{L}_1}(n^2))),$$

and

$$H_n = \frac{\mu_S n - \mu_S^2 - \sigma_S^2}{n^2(n-1)v_n \mathrm{e}^{2\mu_S \alpha_n}} \sum_{j=\lfloor \varepsilon n \rfloor}^{k_n} (\rho_n^{2j}(\mathrm{e}^{-\mu_S(j-1)/n} n + o_{\mathcal{L}_1}(n))).$$

For large n, we have

$$\begin{aligned} |C_n| &\le \frac{2(\sigma_S^2 - \mu_S)n}{n^3(n-1)v_n \mathrm{e}^{2\mu_S \alpha_n}} \sum_{j=1}^{\lfloor \varepsilon n \rfloor - 1} \rho_n^{2j}(n^2) \\ &\le \frac{16 s^2}{n v_n \mathrm{e}^{2\mu_S \alpha_n}} \sum_{j=1}^{\lfloor \varepsilon n \rfloor} 2\mathrm{e}^{\mu_S M_1} \\ &= O(\varepsilon) \quad \text{as } \varepsilon \to 0. \end{aligned}$$

Likewise, we have

$$|C_n'| = O(\varepsilon) \quad \text{as } \varepsilon \to 0.$$

The formulae for D_n and H_n involve sums of geometric series. Thus, D_n reduces to

$$\begin{aligned} D_n &= \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{n(n-1)v_n \mathrm{e}^{2\mu_S \alpha_n}} \sum_{j=\lfloor \varepsilon n \rfloor}^{k_n} \rho_n^{2j}(\mathrm{e}^{-2\mu_S j/n} + o_{\mathcal{L}_1}(1)) \\ &= \frac{(\sigma_S^2 - \mu_S)n + \mu_S^2}{n(n-1)v_n \mathrm{e}^{2\mu_S \alpha_n}} \left(\left(\sum_{j=0}^{k_n} \rho_n^{2j} \mathrm{e}^{-2\mu_S j/n} \right) - \left(\sum_{j=0}^{\lfloor \varepsilon n \rfloor - 1} \rho_n^{2j} \mathrm{e}^{-2\mu_S j/n} \right) \right. \\ &\qquad \left. + o_{\mathcal{L}_1}(1) \sum_{j=\lfloor \varepsilon n \rfloor}^{k_n} \rho_n^{2j} \right). \end{aligned}$$

This calculation involves two sums of the form

$$\sum_{j=0}^{b_n - 1} \rho_n^{2j} \mathrm{e}^{-2\mu_S j/n} = \frac{(n/(n-\mu_S))^{2b_n} \mathrm{e}^{-2\mu_S b_n/n} - 1}{(n/(n-\mu_S))^2 \mathrm{e}^{-2\mu_S/n} - 1},$$

with $b_n = \beta_n n + r_n$, and the remainder function r_n is $o(n)$; in one sum β_n is ε, and in the other it is α_n. Using the asymptotic relation

$$\left(\frac{n}{n-\mu_S} \right)^{2\beta_n n} = \mathrm{e}^{2\mu_S \beta_n} + \frac{\mu_S^2 \beta_n \mathrm{e}^{2\mu_S \beta_n}}{n} + O\left(\frac{1}{n^2} \right),$$

and the standard local expansion

$$\mathrm{e}^{c/n} = 1 + \frac{c}{n} + \frac{c^2}{2n^2} + O\left(\frac{1}{n^3} \right),$$

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

we obtain

$$\begin{aligned}\sum_{j=0}^{b_n-1}\rho_n^{2j}\mathrm{e}^{-2\mu_S j/n} &= \left(\left(\mathrm{e}^{2\mu_S\beta_n}+\frac{\mu_S^2\beta_n\mathrm{e}^{2\mu_S\beta_n}}{n}+O\left(\frac{1}{n^2}\right)\right)\right.\\ &\quad\left.\times\left(\frac{n}{n-\mu_S}\right)^{2r_n}\mathrm{e}^{-(2\mu_S\beta_n n+2\mu_S r_n)/n}-1\right)\left(\mu_S^2+O\left(\frac{1}{n}\right)\right)^{-1}(n-\mu_S)^2\\ &= \frac{(1+\mu_S^2\beta_n/n+O(1/n^2))\mathrm{e}^{2r_n(\mu_S/n+O(1/n^2))}\mathrm{e}^{-2\mu_S r_n/n}-1}{\mu_S^2+O(1/n)}(n-\mu_S)^2\\ &= \frac{(1+\mu_S^2\beta_n/n+O(1/n^2))(1+O(r_n/n^2))-1}{\mu_S^2+O(1/n)}(n-\mu_S)^2\\ &= \beta_n n+o(n).\end{aligned}$$

Hence, we have

$$D_n=\frac{1}{v_n\mathrm{e}^{2\mu_S\alpha_n}}((\sigma_S^2-\mu_S)\alpha_n-(\sigma_S^2-\mu_S)\varepsilon)+o(1)+o_{\mathcal{L}_1}(1).$$

Similarly, we have

$$\begin{aligned}\sum_{j=0}^{b_n-1}\rho_n^{2j}\mathrm{e}^{-\mu_S j/n} &= \frac{(n-\mu_S)^2[(n/(n-\mu_S))^{2b_n}\mathrm{e}^{-\mu_S b_n/n}-1]}{n^2\mathrm{e}^{-\mu_S/n}-(n-\mu_S)^2}\\ &= \frac{(n-\mu_S)^2(\mathrm{e}^{2\mu_S\beta_n}+O(1/n))\mathrm{e}^{O(2r_n/n)}\mathrm{e}^{-(\mu_S\beta_n n+o(n))/n}-1}{\mu_S n+O(1)}\\ &= \left(\frac{\mathrm{e}^{\mu_S\beta_n}-1}{\mu_S}\right)n+o(n).\end{aligned}$$

So,

$$H_n=\frac{1}{v_n\mathrm{e}^{2\mu_S\alpha_n}}((\mathrm{e}^{\mu_S\alpha_n}-1)-(\mathrm{e}^{\mu_S\varepsilon}-1))+o(1)+o_{\mathcal{L}_1}(1).$$

Consequently, we have

$$V_n=O(\varepsilon)+\frac{1}{v_n\mathrm{e}^{2\mu_S\alpha_n}}[(\mathrm{e}^{\mu_S\alpha_n}+\alpha_n(\sigma_S^2-\mu_S)-1)-\varepsilon(\mathrm{e}^{\mu_S\varepsilon}+(\sigma_S^2-\mu_S)-1)]+o_{\mathcal{L}_1}(1).$$

Taking the limit as $\varepsilon\to 0$, we obtain

$$\lim_{\varepsilon\to 0}V_n=1+o_{\mathcal{L}_1}(1).$$

Now, let $n\to\infty$ to obtain

$$V_n\xrightarrow{\mathrm{P}}1.$$

Hence, the 1-conditional variance condition is verified.

According to the martingale central limit theorem

$$\sum_{j=1}^{k_n}\left(\frac{\nabla\tilde{Y}_j}{\sqrt{n v_n\mathrm{e}^{2\mu_S\alpha_n}}}\right)=\frac{Y_{k_n}-Y_0}{\sqrt{n v_n\mathrm{e}^{2\mu_S\alpha_n}}}\xrightarrow{\mathrm{D}}\mathcal{N}(0,1).$$

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

Subsequently, we write

$$\frac{\rho_n^{k_n} W_{k_n} - \rho_n^0 n}{\sqrt{n v_n \mathrm{e}^{2\mu_S \alpha_n}}} = \frac{(n/(n-\mu_S))^{k_n} W_{k_n} - n}{\sqrt{n v_n \mathrm{e}^{2\mu_S \alpha_n}}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1).$$

Using the asymptotic relation $(n/(n-\mu_S))^{k_n} = \mathrm{e}^{\mu_S \alpha_n} + O(1/n)$ in the linear phase, it follows that

$$\frac{(\mathrm{e}^{\mu_S \alpha_n} + O(1/n)) W_{k_n} - n}{\sqrt{n v_n \mathrm{e}^{2\mu_S \alpha_n}}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1).$$

However, we have $W_{k_n} \le n$, and $W_{k_n} O(1/n)/\sqrt{n v_n \mathrm{e}^{2\mu_S \alpha_n}} \to 0$; with an application of Slutsky's additive theorem (see Karr (1993, p. 146)), we arrive at

$$\frac{\mathrm{e}^{\mu_S \alpha_n} W_{k_n} - n}{\sqrt{n v_n \mathrm{e}^{2\mu_S \alpha_n}}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1).$$

Now use the relation $R_{k_n} + W_{k_n} = n$ to obtain the theorem as stated.

6.3. The superlinear phase

When the number k_n of draws becomes superlinear, the mean number of uncollected coupons becomes 0 at a very fast rate. Various asymptotic forms of the mean and variance appear depending on the degree of superlinearity. For example, for fixed $S = s$, if $k_n \sim n \ln^2 n$, both the mean and variance of the number of uncollected coupons (white balls) are $n\mathrm{e}^{-2\ln^2 n}$, whereas when $k_n \sim n^2$, both the mean and variance diminish as fast as $\mathrm{e}^{-2n+n^{-1}\ln n+2}$. To obtain a sense for what happens in the superlinear phase, we will only discuss the case of fixed $S = s$. As we show next, with a fixed number of purchases, in general, in the superlinear phase the mean and variance are asymptotically negligible at an exponential rate.

Proposition 1. *In the superlinear phase with a fixed number of purchases,*

$$\mathrm{var}[W_{k_n}] \le \mathrm{E}[W_{k_n}] = n\mathrm{e}^{-\Theta(k_n/n)},$$

where $\Theta(k(n)/n)$ is a positive function of the exact order k_n/n.

Proof. According to the exact variance formula (6), we have

$$\begin{aligned}
\mathrm{var}[W_{k_n}] &\le \left(\frac{n-s}{n}\right)^{k_n} \left(\frac{n-s-1}{n-1}\right)^{k_n} n^2 + \left(\frac{n-s}{n}\right)^{k_n} n - \left(\frac{n-s}{n}\right)^{2k_n} n^2 \\
&\le \left(\frac{n-s}{n}\right)^{k_n} n \\
&= \mathrm{E}[W_{k_n}] \\
&= n\mathrm{e}^{k_n \ln((n-s)/n)} \\
&= n\mathrm{e}^{-\Theta(k_n/n)}.
\end{aligned}$$

Proposition 2. *In the superlinear phase with a fixed number of purchases,*

$$R_{k_n}/n \xrightarrow{\mathrm{P}} 1.$$

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

Proof. By Proposition 1 and Chebyshev's inequality,

$$\begin{aligned}\mathrm{P}\left(\left|\frac{W_{k_n}}{n}-\frac{\mathrm{E}[W_{k_n}]}{n}\right|>\varepsilon\right) &\le \frac{\mathrm{var}[W_{k_n}]}{\varepsilon^2 n^2}\\ &= O\left(\frac{1}{n\mathrm{e}^{\Theta(k_n/n)}}\right)\\ &\to 0 \quad \text{as } n\to\infty.\end{aligned}$$

Hence,

$$\frac{W_{k_n}}{n}-\frac{\mathrm{E}[W_{k_n}]}{n}\xrightarrow{\mathrm{P}}0,$$

and in the superlinear phase $\mathrm{E}[W_{k_n}]/n\to 0$. An application of Slutsky's additive theorem (Karr (1993, p. 146)) yields the result.

7. Concluding remarks

We identified Gaussian phases in the long-term drawing of samples of small size S (independent and identically distributed on $\{1,2,\ldots,n\}$) from an urn representing the generalized coupon problem. We used k_n to denote the number of draws, and W_{k_n} and R_{k_n} to respectively denote the numbers of uncollected and collected coupons at the end of k_n draws. In the entire sublinear phase, we have W_j/n, the proportion of the remaining uncollected coupons, convergent to 1 almost surely, for any $0\le j\le k_n$. For a genuinely random S (with positive variance), across the entire sublinear phase we have a Gaussian distribution for a suitably shifted and scaled number of coupons collected under general assumptions on the variability of S. The argument breaks down, however, when S is deterministic (whether it grows with n or not) at $k_n\sim\sqrt{n}$. Indeed, if $S=s$ is fixed, Mikhaĭlov (1977) found a Poisson limit for $R_{k_n}-s_nk_n$, when $s_nk_n\sim b\sqrt{n}$ for $b>0$. For deterministic cases, Gaussianity holds only in the upper sublinear phase, where the number of draws $k_n\to\infty$ at a rate higher than $\sqrt{n}$ (but slower than n). By similar methods, if $S=s_n$ is a growing function of n, we can prove Gaussian limit laws only under additional mild conditions, such as $s_n^2k_n^2=o(n)$. Another extension that can be handled by these methods is the case where the purchases are independent but not necessarily identically distributed.

When the number of draws k_n grows linearly, Theorem 3 applies. We mentioned in Subsection 6.2 that the result in the sublinear phase is similar to that in the linear phase. In fact, the result is the same for fixed $S=s$ and pure linearity (no oscillation in the leading term of k_n) under the interpretation that $(\mathrm{e}^{s\alpha}-s\alpha-1)/(\alpha^2\mathrm{e}^{2s\alpha})\to\frac{1}{2}s^2$ as $\alpha\to 0$. Though the details of their proofs are somewhat different, Theorems 1 and 3 can both be viewed as specialized cases of one combined master theorem of the form

$$\frac{R_{k_n}-(1-s/n)^{k_n}n}{k_n/\sqrt{n}}\xrightarrow{\mathrm{D}}\mathcal{N}\left(0,\lim_{a\to\alpha}\frac{\mathrm{e}^{sa}-sa-1}{a^2\mathrm{e}^{2sa}}\right)$$

for $\alpha\ge 0$, which explains what happens at the seam line between the very high end of the sublinear phase with fixed purchases (say when $k_n=\lceil 3000n/\ln n\rceil$), and the very low end of the linear phase (say when $k_n=\lfloor 0.000\,001\,94n-20\sqrt{n}\rfloor$), where in both phases the number of white balls in the urn is asymptotically normal with mean of about n and variance of about $\frac{1}{2}s^2n$.

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

Note that in Theorem 3 if $k_n \sim \alpha_n n$, and the coefficient α_n is not convergent to a limit α, the random variable $(R_{k_n} - (1 - \mathrm{e}^{-s\alpha_n})n)/\sqrt{n}$ does not converge at all. For example, if

$$k_n = \lfloor (5 + 2\cos(\pi n))n + 3n^{0.1} \rfloor,$$

the coefficient of linearity contains a sinusoid that is -1 infinitely often, and $+1$ infinitely often. Therefore, there will be subsequences of $(R_{k_n} - (1 - \mathrm{e}^{-s\alpha_n})n)/\sqrt{n}$ converging to $\mathcal{N}(0, (\mathrm{e}^{3s} - 3s - 1)\mathrm{e}^{-6s})$, and others converging to $\mathcal{N}(0, (\mathrm{e}^{7s} - 7s - 1)\mathrm{e}^{-14s})$, and $(R_{k_n} - (1 - \mathrm{e}^{-s\alpha_n})n)/\sqrt{n}$ does not converge in distribution to any limit. It is only when $(\mathrm{e}^{s\alpha_n} - s\alpha_n - 1)\mathrm{e}^{-2s\alpha_n}$ is subsumed in the scale that we have the convergence

$$\frac{R_{k_n} - (1 - \mathrm{e}^{-s\alpha})n}{\sqrt{n(\mathrm{e}^{s\alpha_n} - s\alpha_n - 1)\mathrm{e}^{-2s\alpha_n}}} \xrightarrow{\mathrm{D}} \mathcal{N}(0, 1).$$

The result of Proposition 2 asserts that, with high probability, almost all the coupons are collected in the superlinear phase, as does Theorem 2 for the linear phase for very large α_n (tending to $+\infty$). This again explains what happens at the seam line between the very high end of the linear phase (say when $k_n = 10\,000\,000n$) and the very low end of the superlinear phase (say when $k_n = \lceil n \ln n - 3n \rceil$).

The result of Proposition 2 applies as soon as k_n becomes superlinear, even when it barely enters that phase, such as in the case when $k_n = \lfloor n \ln \ln \ln n \rfloor$. However, when k_n gets deeper in the superlinear phase, such as in the case $k_n = \lfloor \frac{1}{2} n \ln n \rfloor$, $k_n = \lceil n\sqrt{n} - 2 \rceil$, or $k_n = n^2 + 3n + 6$, the rate of convergence in the sequence of probabilities

$$\mathrm{P}\left(\left| \frac{W_{k_n}}{n} - \frac{\mathrm{E}[W_{k_n}]}{n} \right| > \varepsilon \right) = O\left(\frac{1}{n\mathrm{e}^{\Theta(k_n/n)}} \right)$$

is fast enough to admit the relation

$$\sum_{n=1}^{\infty} \mathrm{P}\left(\left| \frac{W_{k_n}}{n} - \frac{\mathrm{E}[W_{k_n}]}{n} \right| > \varepsilon \right) < \infty,$$

which enables the Borel–Cantelli lemma to hold and give the stronger statement $(W_{k_n} - \mathrm{E}[W_{k_n}])/n \xrightarrow{\text{a.s.}} 0$, or, equivalently, $R_{k_n}/n \xrightarrow{\text{a.s.}} 1$ (as $\mathrm{E}[W_{k_n}]/n \to 0$ in this phase). Smythe (2009) took up the investigation of the asymptotic distributional forms in the superlinear phase.

Other interpretations of coupon collection can be found in the literature of graph theory and occupancy problems. For example, starting with n isolated vertices, a random hypergraph can be generated by adding $k = k_n$ (hyper) edges (which are subsets of vertices, of size s each) and the edges are chosen independently and uniformly at random. This coincides with the case of generalized coupon collection with fixed purchases, $S = s$, at each stage. The uncollected coupons are the vertices that remain isolated in this hypergraph model. Bender *et al.* (1997) provided an enumerative study of these hypergraphs.

In the area of occupancy problems, balls are dropped in urns and one asks questions about empty urns. Starting with n urns, and choosing S urns at a time, we drop S balls (one ball in each chosen urn), we obtain an occupancy problem similar to the coupon collection problem we considered. Urns in this occupancy problem are coupons in our model. Mikhaĭlov (1977), (1980) and Vatutin and Mikhaĭlov (1982) carried out extensive studies for deterministic S. The book of Kolchin *et al.* (1978) thoroughly addresses many variations of occupancy problems with $S \equiv 1$.

https://doi.org/10.1239/aap/1293113148 Published online by Cambridge University Press

Acknowledgments

The author is grateful to Professor Bradley Johnson for an encouraging preliminary discussion and for sharing his manuscript. Professor Brendan McKay provided valuable advice and pointed out several related areas of research and references. Professor Robert Smythe helped the author reach an accurate version of Theorem 1.

References

Adler, I. and Ross, S. M. (2001). The coupon subset collection problem. *J. Appl. Prob.* **38,** 737–746.

Bender, E. A., Canfield, E. R. and McKay, B. D. (1997). The asymptotic number of labeled graphs with n vertices, q edges, and no isolated vertices. *J. Combinatorial Theory A* **80,** 124–150.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Applications*. Academic Press, New York.

Ivchenko, G. I. (1998). How many samples does it take to see all of the balls in an urn? *Math. Notes* **64,** 49–54.

Johnson, B. C. and Sellke, T. M. (2010). On the number of i.i.d. samples required to observe all of the balls in an urn. *Methodology Comput. Appl. Prob.* **12,** 139–154.

Karr, A. F. (1993). *Probability*. Springer, New York.

Kobza, J. E., Jacobson, S. H. and Vaughan, D. E. (2007). A survey of the coupon collector's problem with random sample sizes. *Methodology Comput. Appl. Prob.* **9,** 573–584.

Kolchin, V. F., Sevastyanov, B. A. and Chistyakov, V. P. (1978). *Random Allocations*. John Wiley, New York.

Mikhaĭlov, V. G. (1977). A Poisson limit theorem in the scheme of group disposal of particles. *Theory Prob. Appl.* **22,** 152–156.

Mikhaĭlov, V. G. (1980). Asymptotic normality of the number of empty cells for group allocation of particles. *Theory Prob. Appl.* **25,** 82–90.

Pólya, G. (1930). Eine Wahrscheinlichkeitsaufgabe zur Kundenwerbung. *Z. Angew. Math. Mech.* **10,** 96–97.

Sellke, T. M. (1995). How many i.i.d. samples does it take to see all the balls in a box? *Ann. Appl. Prob.* **5,** 294–309.

Smythe, R. (2009). Phases in generalized coupon collection. Personal communication.

Stadje, W. (1990). The collector's problem with group drawings. *Adv. Appl. Prob.* **22,** 866–882.

Vatutin, V. A. and Mikhaĭlov, V. G. (1982). Limit theorems for the number of empty cells in an equiprobable scheme for group allocation of particles. *Theory Prob. Appl.* **27,** 734–743.