Electronic Journal of Probability

Vol. 14 (2009), Paper no. 74, pages 2156–2181.

Journal URL
http://www.math.washington.edu/~ejpecp/

Normal approximation for isolated balls in an urn allocation model

Mathew D. Penrose[†]
Department of Mathematical Sciences,
University of Bath, Bath BA2 7AY,
United Kingdom
m.d.penrose@bath.ac.uk

Abstract

Consider throwing n balls at random into m urns, each ball landing in urn i with probability p_i. Let S be the resulting number of singletons, i.e., urns containing just one ball. We give an error bound for the Kolmogorov distance from the distribution of S to the normal, and estimates on its variance. These show that if n, m and $(p_i, 1 \leq i \leq m)$ vary in such a way that $\sup_i p_i = O(n^{-1})$, then S satisfies a CLT if and only if $n^2 \sum_i p_i^2$ tends to infinity, and demonstrate an optimal rate of convergence in the CLT in this case. In the uniform case ($p_i \equiv m^{-1}$) with m and n growing proportionately, we provide bounds with better asymptotic constants. The proof of the error bounds is based on Stein's method via size-biased coupling.

Key words: Berry-Esseen bound, central limit theorem, occupancy scheme, size biased coupling, Stein's method.

AMS 2000 Subject Classification: Primary 60F05; Secondary: 62E17, 60C05.

Submitted to EJP on April 8, 2009, final version accepted September 18, 2009.

[†]Partly supported by the Alexander von Humboldt Foundation though a Friedrich Wilhelm Bessel Research Award.

1 Introduction

Consider the classical occupancy scheme, in which each of n balls is placed independently at random in one of m urns, with probability p_i of going into the ith urn ($p_1 + p_2 + \cdots + p_m = 1$). If N_i denotes the number of balls placed in the ith urn, then $(N_1, \ldots, N_m)$ has the multinomial distribution $\mathrm{Mult}(n; p_1, p_2, \ldots, p_m)$. A special case of interest is the so-called uniform case where all the p_i are equal to $1/m$.

A much-studied quantity is the number of occupied urns, i.e. the sum $\sum_i \mathbf{1}\{N_i > 0\}$. This quantity, scaled and centred, is known to be asymptotically normal as $n \to \infty$ in the uniform case with $m \propto n$, and a Berry-Esseen bound for the discrepancy from the normal, tending to zero at the optimum rate, was obtained for the uniform case by Englund [4] (see also [15]), and for the general (nonuniform) case, with a less explicit error bound, by Quine and Robinson [12]. More recently, Hwang and Janson [8] have obtained a local limit theorem. A variety of applications are mentioned in [8] ('coupon collector's problem, species trapping, birthday paradox, polynomial factorization, statistical linguistics, memory allocation, statistical physics, hashing schemes and so on'), and further applications of the occupancy scheme are listed in Feller (Section 1.2 of [5]) and in Gnedin et al. [6]. Also noteworthy are the monographs by Johnson and Kotz [9] and by Kolchin *et al.* [10]; the latter is mainly concerned with models of this type, giving results for a variety of limiting regimes for the growth of m with n (in the uniform case) and also in some of the non-uniform cases. There has also been recent interest in the case of infinitely many urns with the probabilities p_i independent of n [1; 6].

In this paper we consider the number of isolated balls, that is, the sum $\sum_i \mathbf{1}\{N_i = 1\}$. This quantity seems just as natural an object of study as the number of occupied urns, if one thinks of the model in terms of the balls rather than in terms of the urns. For example, in the well-known birthday paradox, this quantity represents the number of individuals in the group who have a unique birthday.

We mention some other applications where the number of isolated balls could be of particular interest. If the balls represent individuals and the urns represent their classification in a database according to certain characteristics (see [6]), then the number of isolated balls represents the number of individuals which can be identified uniquely from their classification. If the balls represent particles or biological individuals, and the urns represent their spatial locations, each particle occupying one of m locations chosen at random, and if two or more particles sharing the same location annihilate one another, then the number of isolated balls is the number of surviving particles. If the balls represents the users of a set of communications channels at a given instant (these could be physical channels or wavebands for wireless communications), and each user randomly selects one of m available channels, and two or more attempted uses of the same channel interfere with each other and are unsuccessful, then the number of isolated balls is the number of successful users.

In the uniform case, we obtain an explicit Berry-Esseen bound for the discrepancy of the number of isolated balls from the normal, tending to zero at the optimum rate when $m \propto n$. In the non-uniform case we obtain a similar result with a larger constant, also giving upper and lower bounds which show that the variance of the number of isolated balls is $\Theta(n^2 \sum_i p_i^2)$. Together, these results for the non-uniform case show that provided $n \max_i \{p_i\}$ is bounded, a central limit theorem for the number of isolated balls holds if and only if $n^2 \sum_i p_i^2$ tends to infinity, and give the rate of convergence to the normal for this case. The proof of the variance bounds, in Section 5, is based on martingale difference techniques and somewhat separate from the other arguments in the paper.

Our Berry-Esseen results for the number of isolated balls are analogous to the main results of [4] (in the uniform case) and [12] (in the non-uniform case) for the number of occupied urns. Our proofs, however, are entirely different. We adapt a method used recently by Goldstein and Penrose [7] for a problem in stochastic geometry (Theorem 2.1 of [7]). A more distantly related method has been used previously for *Poisson* approximation in the occupancy scheme; see Barbour *et al.* [2], pages 11 and 115.

Our method does not involve either characteristic functions, or first Poissonizing the total number of balls; in this, it differs from most of the approaches adopted in the past for normal approximation in occupancy-type schemes. As remarked in [8] 'almost all previous approaches rely, explicitly or implicitly, on the widely used Poissonization technique', and this remark also applies to [8] itself. One exception is Chatterjee [3], who uses a method not involving Poissonization to give an error bound with the optimal rate of decay (with unspecified constant) for the Kantorovich-Wasserstein distance (rather than the Kolmogorov distance, as here) between the distribution of the number of occupied urns and the normal, in the uniform case.

We believe that our approach can be adapted to the number of urns containing k balls, for arbitrary fixed k, but these might require significant amounts of extra work, so we restrict ourselves here to the case with $k = 1$.

Our approach is based on size-biased couplings. Given a nonnegative random variable W with finite mean $\mu = EW$, we say W' has the W size biased distribution if $P[W' \in dw] = (w/\mu)P[W \in dw]$, or more formally, if

$$\mathbb{E}[Wf(W)] = \mu \mathbb{E} f(W') \quad \text{for bounded continuous functions } f. \tag{1.1}$$

Lemma 3.1 below tells us that if one can find coupled realizations of W and W' which are in some sense close, then one may be able to find a good Berry-Esseen bound for W. It turns out that this can be done for the number of non-isolated balls.

2 Results

Let $n \in \mathbb{N}$ and $m = m(n) \in \mathbb{N}$ with $m \geq 4$. Let $\mathbf{p}^{(n)} = (p_x^{(n)}, 1 \leq x \leq m)$ be a probability mass function on $[m] := \{1, 2, \dots, m\}$, with $p_x^{(n)} > 0$ for all $x \in [m]$. Let X and $X_i, 1 \leq i \leq n$ be independent and identically distributed random variables with probability mass function $\mathbf{p} = \mathbf{p}^{(n)}$ (we shall often suppress the superscript (n)). Define $Y = Y(n)$ by

$$M_i := -1 + \sum_{j=1}^{n} \mathbf{1}\{X_j = X_i\}; \quad Y := \sum_{i=1}^{n} \mathbf{1}\{M_i > 0\}. \tag{2.1}$$

In terms of the urn scheme described in Section 1, the probability of landing in Urn x is p_x for each ball, X_i represents the location of the ith ball, M_i represents the number of other balls located in the same urn as the ith ball, and Y represents the number of non-isolated balls, where a ball is said to be isolated if no other ball is placed in the same urn as it is. Thus $n - Y$ is the number of isolated balls, or in other words, the number of urns which contain a single ball.

Let Z denote a standard normal random variable, and let $\Phi(t) := P[Z \leq t] = (2\pi)^{-1/2} \int_{-\infty}^{t} \exp(-x^2/2)dx$. Given any random variable W with finite mean μ_W and standard deviation σ_W satisfying $0 < \sigma_W < \infty$, define

$$D_W := \sup_{t \in \mathbb{R}} \left| P\left[\frac{W - \mu_W}{\sigma_W} \le t \right] - \Phi(t) \right|,$$

the so-called Kolmogorov distance between the distribution of W and the normal. We are concerned with estimating D_Y (which is equal to D_{n-Y}).

We refer to the case where $p_x = m^{-1}$ for each $x \in [m]$ as *the uniform case*. Our main result for the uniform case provides a normal approximation error bound for Y, which is explicit modulo computation of μ_Y and σ_Y, and goes as follows.

Theorem 2.1. *In the uniform case, if $\sigma_Y^3 \ge 24\mu_Y$, then*

$$\begin{aligned} D_Y &\le \left(\frac{\mu_Y}{5\sigma_Y^2}\right)\left(\sqrt{\frac{44}{\sigma_Y} + 5\sqrt{\eta(n,m)}} + \frac{4}{\sqrt{\sigma_Y}}\right)^2 && (2.2) \\ &\le \left(\frac{\mu_Y}{\sigma_Y^2}\right)\left(\frac{24}{\sigma_Y} + 2\sqrt{\eta(n,m)}\right) && (2.3) \end{aligned}$$

with

$$\eta(n,m) := \frac{16}{n} + \frac{4}{n(n-1)} + \frac{24}{m}\left(2 + \frac{n}{m-3} + \frac{n}{m}\right) \tag{2.4}$$

For asymptotics in the uniform case, we allow $m = m(n)$ to vary with n. We concentrate on the case where $m = \Theta(n)$. (recall that $a_n = \Theta(b_n)$ means a_n/b_n is bounded away from zero and infinity). In this case both μ_Y and σ_Y^2 turn out to be $\Theta(n)$ as $n \to \infty$, and thus Theorem 2.1 implies D_Y is $O(n^{-1/2})$ in this regime. More formally, we have the following.

Theorem 2.2. *Suppose n, m both go to infinity in a linked manner, in such a way that $n/m \to \alpha \in (0,\infty)$. Then with $g(\alpha) := (e^{-\alpha} - e^{-2\alpha}(\alpha^2 - \alpha + 1))^{1/2}$, we have in the uniform case that $g(\alpha) > 0$ and*

$$\begin{aligned} \limsup_{n\to\infty} n^{1/2} D_Y &\le \left(\frac{1-e^{-\alpha}}{5g(\alpha)^2}\right)\left(\sqrt{\frac{44}{g(\alpha)} + 20\sqrt{1+3\alpha(1+\alpha)}} + \frac{4}{\sqrt{g(\alpha)}}\right)^2 && (2.5) \\ &\le \left(\frac{1-e^{-\alpha}}{g(\alpha)^2}\right)\left(\frac{24}{g(\alpha)} + 8\sqrt{1+3\alpha(1+\alpha)}\right). && (2.6) \end{aligned}$$

In the case $\alpha = 1$, the right hand side of (2.5), rounded up to the nearest integer, comes to 172. Theorems 2.1 and 2.2 are proved in Section 4.

We now state our results for the general (non-uniform) case. Given n we define the parameters

$$\|\mathbf{p}\| := \sup_{x \in [m]} (p_x); \qquad \gamma = \gamma(n) := \max(n\|\mathbf{p}\|, 1). \tag{2.7}$$

For the large-n asymptotics we essentially assume that $\gamma(n)$ remains bounded, or at least grows only slowly with n; see Corollary 2.1 below. First we give a non-asymptotic result.

Theorem 2.3. *It is the case that*

$$D_Y \geq \min\left(1/6, (8\pi e)^{-1/2}\sigma^{-1}\right), \tag{2.8}$$

and if

$$\|\mathbf{p}\| \leq 1/11 \tag{2.9}$$

and also

$$n \geq 83\gamma^2(1+3\gamma+3\gamma^2)e^{1.05\gamma}, \tag{2.10}$$

then

$$D_Y \leq 1633\gamma^2 e^{2.1\gamma}\left(\sqrt{99+5C(\gamma)}+6\right)^2 \sigma_Y^{-1}, \tag{2.11}$$

with

$$C(\gamma) := 10(82\gamma^7+82\gamma^6+80\gamma^5+47\gamma^4+12\gamma^3+12\gamma^2)^{1/2}. \tag{2.12}$$

It is of use in proving Theorem 2.3, and also of independent interest, to estimate the variance σ_Y^2 in terms of the original parameters $(p_x, x \in [m])$, and our next result does this. Throughout, we write $\sum_x$ for $\sum_{x=1}^m$.

Theorem 2.4. *It is the case that*

$$\mathrm{Var} Y \leq 8n^2 \sum_x p_x^2, \tag{2.13}$$

and if (*2.9*) *and* (*2.10*) *hold, then*

$$\mathrm{Var} Y \geq (7776)^{-1}\gamma^{-2}e^{-2.1\gamma}n^2\sum_x p_x^2. \tag{2.14}$$

If $\gamma(n)$ remains bounded, i.e. $\sup_n \gamma(n) < \infty$. then both (2.9) and (2.10) hold for large enough n. Hence, the following asymptotic result is immediate from Theorems 2.3 and 2.4. This result gives us, for cases with $\gamma(n)$ bounded, necessary and sufficient conditions in terms of the parameters p_x for a central limit theorem to hold, and gives the rate of convergence of D_Y to zero when it does hold.

Corollary 2.1. *Suppose* $\sup_n \gamma(n) < \infty$*. Then the following three conditions are equivalent:*

- $n^2 \sum_x p_x^2 \to \infty$ *as* $n \to \infty$ *;*
- $\sigma_Y \to \infty$ *as* $n \to \infty$ *;*
- $(Y - \mathbb{E}Y)/\sigma_Y$ *converges in distribution to* Z *as* $n \to \infty$*.*

If these conditions hold, then

$$D_Y = \Theta(\sigma_Y^{-1}) = \Theta\left(\left(n^2\sum_x p_x^2\right)^{-1/2}\right).$$

Remarks. In the uniform case, Theorem 2.2 provides an alternative proof of the central limit theorem for Y when $m = \Theta(n)$ (see Theorem II.2.4 on page 59 of [10]), with error bounds converging to zero at the optimum rate. Corollary 2.1 shows that in the uniform case, if $n^2/m \to \infty$ and n/m remains bounded, then $D_Y = \Theta((n^2/m)^{-1/2})$. Corollary 2.1 overlaps Theorem III.5.2 on page 147 of [10] but is under weaker conditions than those in [10], and provides error bounds not given in [10]. Mikhailov [11] previously provided error bounds for the non-uniform case which were non-optimal (for example, $O(n^{-1/12})$ when $\sum_x p_x^2 = \Theta(n^{-1})$).

The condition that $\gamma(n)$ remain bounded, in Corollary 2.1, is also required by [12] for the analogous Berry-Esseen type result for the number of occupied boxes, though not by [8] for the local limit theorem for that quantity. In (2.9), which is used for the non-asymptotic bounds, the bound of $\frac{1}{11}$ could be replaced by any constant less than $\frac{1}{3}$ without changing anything except the constants in (2.11) and (2.14).

As always (see remarks in [8], [13], [1], and for the uniform case [15]), it might be possible to obtain similar results to those presented here by other methods. However, to do so appears to be a non-trivial task, particularly in the non-uniform case. In [10] the count of the number of isolated balls is treated separately, and differently, from the count of occupied urns or the count of urns with k balls, $k = 0$ or $k \geq 2$. Poisson approximation methods might be of use in some limiting regimes (see [2], Chapter 6), but not when the ratio between $\mathbb{E}[Y]$ and $\mathrm{Var}[Y]$ remains bounded but is not asymptotically 1, which is typically the case here.

Exact formulae can be written down for the probability mass function and cumulative distribution function of Y. For example, using (5.2) on page 109 of [5], the cumulative distribution of Y may be written as

$$P[Y \leq n-k] = P[n-Y \geq k] = \sum_{j=k}^{m} (-1)^{j-k} \binom{j-1}{k-1} S_j$$

with S_j a sum of probabilities that j of the urns contain one ball each, i.e.

$$S_j = \sum_{x_1 < x_2 < \cdots < x_j \leq m} \frac{n!}{(n-j)!} p_{x_1} \cdots p_{x_j} \left(1 - \sum_{i=1}^{j} p_{x_i}\right)^{n-j}.$$

We shall not use this formula in obtaining our normal approximation results.

3 Lemmas

A key tool in our proofs is the following result.

Lemma 3.1. *Let $W \geq 0$ be a random variable with mean μ and variance $\sigma^2 \in (0,\infty)$, and let W^s be defined on the same sample space, with the W-size biased distribution. If $|W^s - W| \leq B$ for some $B \leq \sigma^{3/2}/\sqrt{6\mu}$, then*

$$D_W \leq \frac{\mu}{5\sigma^2}\left(\sqrt{\frac{11B^2}{\sigma} + 5\Delta} + \frac{2B}{\sqrt{\sigma}}\right)^2 \quad (3.1)$$

$$\leq \frac{2\mu}{\sigma^2}\left(\frac{3B^2}{\sigma} + \Delta\right) \quad (3.2)$$

where

$$\Delta := \sqrt{\mathrm{Var}(\mathbb{E}(W^s - W|W))}. \tag{3.3}$$

Proof. The bound (3.1) is given as Lemma 3.1 of [7], where it is proved via Stein's method. The simpler bound (3.2) follows by applying the inequality $x^2+y^2 \leq 2(x^2+y^2)$, valid for all real x, y. □

Our next lemma is concerned with the construction of variables with size-biased distributions.

Lemma 3.2. *Suppose Y is a random variable given by $Y = aP[A|\mathscr{F}]$, where $\mathscr{F}$ is some σ-algebra, $a > 0$ is a constant, and A is an event with $0 < P[A] < 1$. Then Y' has the Y size biased distribution if*

$$\mathscr{L}(Y') = \mathscr{L}(Y|A).$$

Proof. See Lemma 4.1 of [7]. □

Let $\mathrm{Bin}(n,p)$ denote the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in (0,1)$. The following lemma will be used for constructing the desired close coupling of our variable of interest Y, and its size biased version Y', so as to be able to use Lemma 3.1.

Lemma 3.3. *Let $\nu \in \mathbb{N}$ and $p \in (0,1)$. Suppose $N \sim \mathrm{Bin}(\nu,p)$. Define π_k for $k = 0,1,2,\dots,\nu$, by*

$$\pi_k := \begin{cases} \frac{P[N>k|N>0]-P[N>k]}{P[N=k](1-(k/\nu))} & \text{if} \quad 0 \leq k \leq \nu-1 \\ 0 & \text{if} \quad k = \nu. \end{cases} \tag{3.4}$$

Then $0 \leq \pi_k \leq 1$ for each $k \in \{0,1,\dots,\nu\}$.

Proof. See Lemma 3.3 of [7]. □

Our next lemma is a bound on correlations between variables associated with different balls in the urn model. Recall the definition of M_i at (2.1)

Lemma 3.4. *Let $k \in \mathbb{N}$ with $2 \leq k \leq n$, and suppose that for $i = 1,\dots,k$, ψ_i is a real-valued function defined on $\{0\} \cup [n-1]$, with $\mathbb{E}[\psi_1(M_1)] = 0$, and set $\|\psi_i\| := \sup_{\ell\in[n]}\{|\psi_i(\ell-1)|\}$ and $\mathrm{rng}(\psi_i) := \sup_{\ell\in[n]}\{\psi_i(\ell-1)\} - \inf_{k\in[n]}\{\psi_i(\ell-1)\}$. Then in the uniform case,*

$$\left|\mathbb{E}\left[\prod_{i=1}^{k}\psi_i(M_i)\right]\right| \leq \frac{k-1}{m}\left(\prod_{i=2}^{k}\|\psi_i\|\right)\mathrm{rng}(\psi_1)\left(2+\left(\frac{n}{m-k+1}\right)+\frac{n}{m}\right).$$

Proof. Set $W := \prod_{i=1}^{k}\psi_i(M_i)$. Write $\mathbf{X}^{(k)}$ for $(X_1,\dots,X_k)$ and $\mathbf{x}$ for k-vectors $\mathbf{x} = (x_1,\dots,x_k)$ of possible values of $\mathbf{X}^{(k)}$. Let F be the set of vectors $\mathbf{x} = (x_1,\dots,x_k)$ such that $x_1 \neq x_j$ for $j = 2,\dots,k$, so that $\{\mathbf{X}^{(k)} \in F\}$ is the event that each of Balls $2,\dots,k$ lands in a different urn from Ball 1. Then $P[\mathbf{X}^{(k)} \in F^c] \leq (k-1)/m$, and $|W| \leq \prod_{i=1}^{k}\|\psi_i\|$, so that

$$\begin{aligned}|\mathbb{E}\,W| &= \left|E[W|\mathbf{X}^{(k)} \in F^c]P[\mathbf{X}^{(k)} \in F^c] + \sum_{\mathbf{x}\in F}P[\mathbf{X}^{(k)} = \mathbf{x}]\mathbb{E}[W|\mathbf{X}^{(k)} = \mathbf{x}]\right| \\ &\leq ((k-1)/m)\prod_{i=1}^{k}\|\psi_i\| + \sup_{\mathbf{x}\in F}|\mathbb{E}[W|\mathbf{X}^{(k)} = \mathbf{x}]|. \end{aligned}\tag{3.5}$$

Fix $\mathbf{x} \in F$. We group the urns into three 'boxes'. Let Box 1 consist of the urn containing Ball 1, and let Box 2 be the union of the urns containing Balls $2, 3, \ldots, k$; this could be the union of any number up to $k-1$ of urns depending on how many of $x_2, x_3, \ldots, x_k$ are distinct, but since we assume $\mathbf{x} \in F$, Box 2 does not overlap Box 1. Let Box 3 consist of all other urns except those in Box 1 or Box 2. For $i = 1, 2, 3$, let N_i be the number of balls in Box i, other than Balls $1, \ldots, k$. Let $h(\ell)$ be the expected value of $\prod_{i=2}^{k} \psi_i(M_i)$, given $\mathbf{X}^{(k)} = \mathbf{x}$ and given that $N_2 = \ell$. Let $P_{\mathbf{x}}$ and $\mathbb{E}_{\mathbf{x}}$ denote conditional probability and conditional expectation, given that $\mathbf{X}^{(k)} = \mathbf{x}$. Then

$$\begin{aligned} \mathbb{E}_{\mathbf{x}}[W] &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\mathbf{x}}[W|N_1, N_2]] \\ &= \mathbb{E}_{\mathbf{x}}\left[\psi_1(N_1)\mathbb{E}_{\mathbf{x}}\left[\prod_{i=2}^{k} \psi_i(M_i)|N_1, N_2\right]\right] \\ &= \mathbb{E}_{\mathbf{x}}[\psi_1(N_1)h(N_2)] \end{aligned} \tag{3.6}$$

where the last line follows because $\mathbb{E}_{\mathbf{x}}\left[\prod_{i=2}^{k} \psi_i(M_i)|N_1, N_2\right]$ depends on (N_1, N_2) only through N_2. Also, given $\mathbf{X}^{(k)} = \mathbf{x}$, $(N_i)_{i=1}^{3}$ have the multinomial distribution

$$(N_1, N_2, N_3) \sim \text{Mult}\left(n-k; \frac{1}{m}, \frac{a}{m}, 1 - \frac{1+a}{m}\right), \tag{3.7}$$

and a denotes the number of distinct values taken by $x_2, \ldots, x_k$.

We give a coupling of N_1 to another random variable N_1' with the same distribution as M_1 that is independent of N_2, for which we can give a useful bound on $P[N_1 \neq N_1']$. The idea here is that the distribution of N_1, given $N_2 = j$, is obtained from throwing $n-k-j$ balls conditioned to avoid Box 2; this conditioning can be effected by re-throwing those balls which land in Box 2. Hence N_1 with this conditioned distribution, coupled to N_1' with the unconditioned distribution of M_1 can be obtained by first throwing $n-k-j$ balls, then throwing $j+k-1$ extra balls to get N_1', and re-throwing the balls which landed in Box 2 (and ignoring the extra balls) to get N_1.

To give this coupling in more detail, we imagine a colouring scheme. Consider throwing a series of coloured balls so each ball can land in one of the three boxes, where the probabilities of landing in Boxes $1, 2, 3$ are $1/m$, a/m, $(m-a-1)/m$ respectively. First, throw $n-k$ white balls and let N_1^*, N_2, N_3^* be the number of white balls in Boxes $1, 2, 3$ respectively. Then pick out the balls in Boxes 1 and 3, paint them red, and throw them again. Then throw enough green balls so the total number of green and red balls is $n-1$. Finally take the red balls in Box 2 (of which there are N_0, say), paint them blue, and throw them again but condition them to land in Boxes 1 and 3 (or equivalently, throw each blue ball again and again until it avoids Box 2). Then (with obvious notation, superscripts denoting colours) set

$$N_1 = N_1^r + N_1^b, \quad N_3 = N_3^r + N_3^b, \quad N_1' = N_1^r + N_1^g.$$

Then (N_1, N_2, N_3) have the multinomial distribution given by (3.7). Also, N_1' has the same $\text{Bin}(n-1, \frac{1}{m})$ as M_1 in the statement of the lemma, and N_1' is independent of N_2. Since $N_1' = N_1 - N_1^b + N_1^g$, we have that

$$\begin{aligned} P_{\mathbf{x}}[N_1 \neq N_1'] &\leq \mathbb{E}_{\mathbf{x}}[N_1^g] + \mathbb{E}_{\mathbf{x}}[N_1^b] \\ &\leq \frac{1}{m}\left(k-1+\mathbb{E}_{\mathbf{x}}N_2\right) + \left(\frac{1/m}{1-(k-1)/m}\right)\mathbb{E}_{\mathbf{x}}[N_0] \\ &\leq \frac{1}{m}\left(k-1+(k-1)n/m + \left(\frac{1}{1-(k-1)/m}\right)(k-1)n/m\right) \end{aligned}$$

so that

$$\left|\mathbb{E}_{\mathbf{x}}[(\psi_1(N_1)-\psi_1(N_1'))h(N_2)]\right| \le P_{\mathbf{x}}[N_1 \ne N_1']\mathrm{rng}(\psi_1)\prod_{i=2}^{k}\|\psi_i\|$$
$$\le \frac{k-1}{m}\left(\left(\frac{m+n}{m}\right)+\left(\frac{n}{m-k+1}\right)\right)\mathrm{rng}(\psi_1)\prod_{i=2}^{4}\|\psi_i\|$$

and since N_1' is independent of N_2 with the same distribution as M_1 (for any $\mathbf{x}$), so that $\mathbb{E}_{\mathbf{x}}\psi_1(N_1') = \mathbb{E}\,\psi_1(M_1) = 0$ by assumption,

$$\mathbb{E}_{\mathbf{x}}[\psi_1(N_1')h(N_2)] = 0,$$

so by (3.6),

$$\left|\mathbb{E}\left[W|\mathbf{X}^{(k)}=\mathbf{x}\right]\right| \le \frac{k-1}{m}\left(\left(\frac{m+n}{m}\right)+\left(\frac{n}{m-k+1}\right)\right)\mathrm{rng}(\psi_1)\prod_{i=2}^{4}\|\psi_i\|.$$

Combined with (3.5), and the fact that $\|\psi_1\| \le \mathrm{rng}(\psi_1)$ since $\mathbb{E}[\psi_1(M_1)] = 0$, this demonstrates the result. $\square$

Next, we adapt Lemma 3.4 to the non-uniform setting. In this case, we need to allow ψ_i to depend on the location as well as the occupation number associated with the ith ball. Consequently, some modification of the proof is required, and the constants in Lemma 3.4 are better than those which would be obtained by simply applying the next lemma to the uniform case.

Lemma 3.5. *Suppose that for $i = 1,2,3,4$, ψ_i is a real-valued function defined on $[m]\times\{0,1,\dots,n-1\}$, with $\mathbb{E}[\psi_1(X_1,M_1)] = 0$, set $\|\psi_i\| := \sup_{(x,k)\in[m]\times[n]}\{|\psi_i(x,k-1)|\}$ and set $\mathrm{rng}(\psi_i) := \sup_{(x,k)\in[m]\times[n]}\{|\psi_i(x,k-1)|\} - \inf_{(x,k)\in[m]\times[n]}\{|\psi_i(x,k-1)|\}$. Assume (2.9) holds. Then*

$$\left|\mathbb{E}\left[\psi_1(X_1,M_1)\psi_2(X_2,M_2)\right]\right| \le (3+3\gamma)\mathrm{rng}(\psi_1)\|\psi_2\|\sum_x p_x^2 \tag{3.8}$$

and

$$\left|\mathbb{E}\left[\prod_{i=1}^{4}\psi_i(X_i,M_i)\right]\right| \le (9+9\gamma)\,\mathrm{rng}(\psi_1)\left(\prod_{i=2}^{4}\|\psi_i\|\right)\sum_x p_x^2. \tag{3.9}$$

Proof. We first prove (3.9). Throw n balls according to the distribution $\mathbf{p}$, with four of them distinguished as Ball 1, Ball 2, Ball 3 and Ball 4. For $i = 2,3,4$, let Z_i be the location of Ball i and let N_i be the number of other balls in the same urn as Ball i. Set $A = \cup_{i=2}^4\{Z_i\}$, the union of the locations of Balls 2,3, and 4.

We shall give a construction of coupled random variables (Z_1,N_1,Z_1',N_1') such that $((Z_i,N_i)_{i=1}^4)$ have the same joint distribution as $((X_i,M_i)_{i=1}^4)$ and (Z_1',N_1') has the same distribution as (X_1,M_1) and is independent of $((Z_i,N_i)_{i=2}^4)$, and such that (Z_1,N_1) is close to (Z_1',N_1') in probability. To do this, given the values of $(Z_i,N_i)_{i=2}^4$ we shall arrange for (Z_1,N_1) to have the conditional distribution of (X_1,M_1) given $(X_i,M_i)_{i=2}^4 = (Z_i,N_i)_{i=2}^4$, while (Z_1',N_1') will have the unconditioned distribution of (X_1,M_1). Similarly to before, this is done, if initially there are $3+j$ balls in A, by first throwing

$n-3-j$ balls, then re-throwing those balls which land in A (to get the conditional distribution) and throwing $3+j$ extra balls (to get the unconditional distribution). Again we explain in more detail via a colouring scheme.

Suppose the balls in A are painted white. Let the balls not in A (including Ball 1 if it is not in A) be re-thrown (again, according to the distribution $\mathbf{p}$). Those which land in A when re-thrown are painted yellow, and the others are painted red.

Now introduce one green ball for each white ball, and if Ball 1 is white, let one of the green balls be labelled Ball G1. Throw the green balls using the same distribution $\mathbf{p}$. Also, introduce a number of blue balls equal to the number of yellow balls, and if Ball 1 is yellow then label one of the blue balls as Ball B1. Throw the blue balls, but condition them to avoid A; that is, use the probability mass function $(p_x/(1-\sum_{y\in A} p_y), x \in [m] \setminus A)$ for the blue balls.

Set Z_1 to be the location of Ball 1 (if it is white or red) or Ball B1 (if Ball 1 is yellow). Set Z_1' to be the location of Ball 1, if it is red or yellow, or the location of Ball G1 (if Ball 1 is white). Let N_1^w, N_1^r, and N_1^b respectively denote the number of white, red, and blue balls at location Z_1, not counting Ball 1 or Ball B1 itself. Let N_1^y, N_1^r, and N_1^g respectively denote the number of yellow, red, and green balls at location Z_1', not counting Ball 1 or Ball G1 itself. Set

$$N_1 = N_1^w + N_1^r + N_1^b, \quad N_1' = N_1^y + N_1^r + N_1^g.$$

Then $((Z_i, N_i)_{i=1}^4)$ have the same joint distribution as $((X_i, M_i)_{i=1}^4)$. Also, (Z_1', N_1') has the same distribution as (X_1, M_1) and (Z_1', N_1') is independent of $((Z_i, N_i)_{i=2}^4)$. Finally, if Ball 1 is red then $Z_1 = Z_1'$ and $N_1' = N_1 - N_1^b + N_1^g$, so that

$$P[(Z_1, N_1) \neq (Z_1', N_1')] \leq \mathbb{E}[N_1^g] + \mathbb{E}[N_1^b] + 2P[Z_1' \in A]. \tag{3.10}$$

Now,

$$P[Z_1' \in A] \leq \sum_{i=2}^{4} P[X_1 = X_i] = 3\sum_x p_x^2. \tag{3.11}$$

Also, if N_g denotes the number of green balls, not including Ball G1 if Ball 1 is green, then by (2.7),

$$\mathbb{E}[N_g] \leq 3 + 3n\|\mathbf{p}\| \leq 3(1+\gamma)$$

and also $\mathbb{E}[N_1^g | N_g] \leq N_g \sum_x p_x^2$, so that

$$\mathbb{E}[N_1^g] = \mathbb{E}[\mathbb{E}[N_1^g | N_g]] \leq 3(1+\gamma)\sum_x p_x^2. \tag{3.12}$$

If N_y denotes the number of yellow balls, other than Ball 1, then by (2.7),

$$\mathbb{E}[N_y] \leq 3n\|\mathbf{p}\| \leq 3\gamma \tag{3.13}$$

and by (2.9),

$$\mathbb{E}[N_1^b | N_y] \leq N_y \sum_x \left(\frac{p_x}{1-3\|\mathbf{p}\|}\right)^2 \leq 2N_y \sum_x p_x^2,$$

so that

$$\mathbb{E}[N_1^b] = \mathbb{E}[\mathbb{E}[N_1^b|N_y]] \le 6\gamma \sum_x p_x^2. \tag{3.14}$$

Set $U := \prod_{i=2}^4 \psi_i(Z_i, N_i)$. By (3.10), (3.11), (3.12) and (3.14),

$$\begin{aligned} \left|\mathbb{E}[U(\psi_1(Z_1,N_1) - \psi_1(Z_1',N_1'))]\right| &\le P[(Z_1,N_1) \ne (Z_1',N_1')]\mathrm{rng}(\psi_1)\prod_{i=2}^4 \|\psi_i\| \\ &\le (9+9\gamma)\mathrm{rng}(\psi_1)\left(\prod_{i=2}^4 \|\psi_i\|\right)\sum_x p_x^2. \end{aligned} \tag{3.15}$$

Since (Z_1', N_1') is independent of $((Z_i, N_i)_{i=2}^4)$ with the same distribution as (X_1, M_1), and $\mathbb{E}\,\psi_1(X_1, M_1) = 0$ by assumption, $\mathbb{E}[U\psi_1(Z_1', N_1')] = 0$. Hence,

$$\mathbb{E}\left[\prod_{i=1}^4 \psi_i(X_i, M_i)\right] = \mathbb{E}\left[U\psi_1(Z_1,N_1)\right] = \mathbb{E}[U(\psi_1(Z_1,N_1) - \psi_1(Z_1',N_1'))],$$

and then (3.9) follows by (3.15). The proof of (3.8) is similar, with the factors of 3 replaced by 1 in (3.11), (3.12) and (3.13). □

4 Proof of Theorems 2.1 and 2.2

Proof of Theorem 2.1. Recall the definition (2.1) of M_i and Y. Assume the uniform case, i.e. assume $\mathbf{p} = (m^{-1}, m^{-1}, \dots, m^{-1})$. Let $\xi_i := \mathbf{1}\{M_i > 0\}$ be the indicator of the event that ball i is not isolated. Then $Y = \sum_{i=1}^n \xi_i$, and we claim that a random variable Y' with the size-biased distribution of Y can be obtained as follows. Let I be a discrete uniform random variable over $[n]$, independent of $X_1, \dots, X_n$. Given the value of I, let $\mathbf{X}' = (X_1', \dots, X_n') \in [m]^n$ be a random n-vector with $\mathscr{L}(X_1', \dots, X_n') = \mathscr{L}(X_1, \dots, X_n | \xi_I = 1)$. Set

$$Y' := \sum_{i=1}^n \mathbf{1}\{\cup_{j \in [n]\setminus\{i\}}\{X_j' = X_i'\}\}.$$

Then $\mathscr{L}(Y') = \mathscr{L}(Y|\xi_I = 1)$ and since $Y = nP[\xi_I = 1|X_1, \dots, X_n]$, the claim follows from Lemma 3.2.

To apply Lemma 3.1 we need to find a random variable Y'', coupled to Y, such that $\mathscr{L}(Y'') = \mathscr{L}(Y')$ and for some constant B we have $|Y'' - Y| \le B$ (almost surely). To check that $\mathscr{L}(Y'') = \mathscr{L}(Y')$, we shall use the fact that if M_I' denotes the number of entries X_j' of $\mathbf{X}'$ that are equal to X_I', other than X_I' itself, then (i) given I and X_I', M_I' has the distribution of a $\mathrm{Bin}(n-1, 1/m)$ variable conditioned to take a non-zero value, and (ii) given I, X_I' and M_I', the distribution of $\mathbf{X}'$ is uniform over all possibilities consistent with the given values of I, X_I' and M_I'.

Define the random n-vector $\mathbf{X} := (X_1 \dots, X_n)$. We can manufacture a random vector $\mathbf{X}'' = (X_1'', \dots, X_n'')$, coupled to $\mathbf{X}$ and (we assert) with the same distribution as $\mathbf{X}'$, as follows.

- Sample the random variables $(X_1, \dots, X_n)$. Define M_i by (2.1).

- Sample a value of I from the discrete uniform distribution on $[n]$, independent of $\mathbf{X}$.
- Sample a Bernoulli random variable $\mathscr{B}$ with $P[\mathscr{B}=1]=\pi_{M_I}$, where $(\pi_k, k \geq 0)$ is given by (3.4) with $\nu = n-1$ and $p = m^{-1}$. (By Lemma 3.3, $0 \leq \pi_k \leq 1$.)
- Sample a value of J from the discrete uniform distribution on $[n] \setminus \{I\}$.
- Define $(X_1'', \dots, X_n'')$ by

$$X_i'' = \begin{cases} X_I & \text{if } i = J \text{ and } \mathscr{B} = 1 \\ X_i & \text{otherwise.} \end{cases}$$

Thus $\mathbf{X}''$ is obtained from $\mathbf{X}$ by changing a randomly selected entry of $\mathbf{X}$ to the value of X_I, if $\mathscr{B}=1$, and leaving $\mathbf{X}$ unchanged if $\mathscr{B}=0$.

We claim that $\mathscr{L}(\mathbf{X}'') = \mathscr{L}(\mathbf{X}')$. To see this define $N := M_I$, and set $N'' := -1 + \sum_{i=1}^n \mathbf{1}\{X_i'' = X_I''\}$. Then N has the $\mathrm{Bin}(n-1, m^{-1})$ distribution, while N'' always takes the value either N or $N+1$, taking the latter value in the case where $\mathscr{B}=1$ and also $X_J \neq X_I$. Thus for any $k \in \{0, 1, \dots, n-1\}$,

$$P[N'' > k] = P[N > k] + P[N = k]\pi_k(1 - (k/(n-1))), \tag{4.1}$$

so by the definition (3.4) of π_k, $\mathscr{L}(N'') = \mathscr{L}(N|N>0)$. This also applies to the conditional distribution of N'' given the values of I and X_I.

Given the values of N'', I and X_I'', the conditional distribution of $\mathbf{X}''$ is uniform over all possibilities consistent with these given values. Hence, $\mathscr{L}(\mathbf{X}'') = \mathscr{L}(\mathbf{X}')$. Therefore setting

$$Y'' := \sum_{i=1}^n \mathbf{1}\{\cup_{j \in [n]\setminus\{i\}} \{X_j'' = X_i''\}\},$$

we have that $\mathscr{L}(Y'') = \mathscr{L}(Y')$, i.e. Y'' has the size-biased distribution of Y.

The definition of $\mathbf{X}''$ in terms of $\mathbf{X}$ ensures that we always have $|Y - Y''| \leq 2$ (with equality if $M_I = M_J = 0$) ; this is explained further in the course of the proof of Proposition 4.1 below. Thus we may apply Lemma 3.1 with $B = 2$. Theorem 2.1 follows from that result, along with the following:

Proposition 4.1. *It is the case that* $\mathrm{Var}(\mathbb{E}[Y'' - Y|Y]) \leq \eta(n,m)$, *where* $\eta(n,m)$ *is given by* (*2.4*).

Proof. Let $\mathscr{G}$ be the σ-algebra generated by $\mathbf{X}$. Then Y is $\mathscr{G}$-measurable. By the conditional variance formula, as in e.g. the proof of Theorem 2.1 of [7],

$$\mathrm{Var}(\mathbb{E}[Y'' - Y|Y]) \leq \mathrm{Var}(\mathbb{E}[Y'' - Y|\mathscr{G}]), \tag{4.2}$$

so it suffices to prove that

$$\mathrm{Var}(\mathbb{E}[Y'' - Y|\mathscr{G}]) \leq \eta(n,m). \tag{4.3}$$

For $1 \leq i \leq n$, let V_i denote the conditional probability that $\mathscr{B}=1$, given $\mathbf{X}$ and given that $I = i$, i.e.

$$V_i = \pi_{M_i}.$$

Let R_{ij} denote the increment in the number of non-isolated balls when the value of X_j is changed to X_i. Then

$$\mathbb{E}[Y'' - Y|\mathcal{G}] = \frac{1}{n(n-1)} \sum_{(i,j):i\neq j} V_i R_{ij}$$

where $\sum_{(i,j):i\neq j}$ denotes summation over pairs of distinct integers i, j in $[n]$.
For $1 \leq i \leq n$ and $j \neq i$, let

$$S_i := \mathbf{1}\{M_i = 0\}; \qquad T_i := \mathbf{1}\{M_i = 0\} - \mathbf{1}\{M_i = 1\};$$
$$Q_{ij} := \mathbf{1}\{M_i = 1\}\mathbf{1}\{X_i = X_j\}.$$

Then we assert that R_{ij}, the increment in the number of non-isolated balls when ball j is moved to the location of ball i, is given by $R_{ij} := S_i + T_j + Q_{ij}$. Indeed, if $X_i \neq X_j$ then S_i is the increment (if any) due to ball i becoming non-isolated, while T_j is the increment (if any) due either to ball j becoming non-isolated, or to another ball at the original location of ball j becoming isolated when ball j is moved to the location of ball i. The definition of Q_{ij} ensures that if $X_i = X_j$ then $S_i + T_j + Q_{ij} = 0$. Thus,

$$\begin{aligned}\mathbb{E}[Y'' - Y|\mathcal{G}] &= \frac{1}{n(n-1)} \sum_{(i,j):i\neq j} V_i(S_i + T_j + Q_{ij}) \\ &= \frac{1}{n}\sum_{i=1}^{n} V_i \tau_i + \frac{1}{n(n-1)} \sum_{(i,j):i\neq j} V_i T_j, \end{aligned} \tag{4.4}$$

where we set

$$\tau_i := S_i + \left(\frac{1}{n-1}\right) \sum_{j:j\neq i} Q_{ij} = \mathbf{1}\{M_i = 0\} + \left(\frac{1}{n-1}\right) \mathbf{1}\{M_i = 1\}. \tag{4.5}$$

Put $a := \mathbb{E}[V_i]$ (this expectation does not depend on i). Then by (4.4),

$$\mathbb{E}[Y'' - Y|\mathcal{G}] = \frac{1}{n}\sum_{i=1}^{n} \left(V_i \tau_i + a T_i\right) + \sum_{(i,j):i\neq j} \frac{(V_i - a)T_j}{n(n-1)}.$$

Since $(x + y)^2 \leq 2(x^2 + y^2)$ for any real x, y, it follows that

$$\begin{aligned}\mathrm{Var}\left(\mathbb{E}[Y'' - Y|\mathcal{G}]\right) &\leq 2\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\left(V_i\tau_i + aT_i\right)\right) \\ &\quad + 2\mathrm{Var} \sum_{(i,j):i\neq j} \frac{(V_i - a)T_j}{n(n-1)}. \end{aligned} \tag{4.6}$$

From the definitions, the following inequalities hold almost surely:

$$-1 \leq T_i \leq 1; \quad 0 \leq V_i \leq 1; \quad 0 \leq \tau_i \leq 1; \tag{4.7}$$

and hence

$$-1 \leq V_i - a \leq 1; \quad -1 \leq (V_i - a)T_j \leq 1; \quad -1 \leq V_i\tau_i + aT_i \leq 2. \tag{4.8}$$

Set $Z_i := V_i\tau_i + aT_i$, and $\bar{Z}_i := Z_i - \mathbb{E}\,Z_i$. By (4.8), $\text{Var}Z_1 \le \mathbb{E}\,Z_1^2 \le 4$. Also by (4.8), we have $|\bar{Z}_i| \le 3$, and $-1 - \mathbb{E}\,Z_i \le \bar{Z}_i \le 2 - \mathbb{E}\,Z_i$. Hence by the case $k = 2$ of Lemma 3.4,

$$\text{Cov}(Z_1, Z_2) = \mathbb{E}\,[\bar{Z}_1\bar{Z}_2] \le \frac{9}{m}\left(2 + \left(\frac{n}{m-1}\right) + \frac{n}{m}\right).$$

Thus for the first term in the right hand side of (4.6), we have

$$\begin{aligned}\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) &= n^{-1}\text{Var}(Z_1) + \left(\frac{n-1}{n}\right)\text{Cov}(Z_1, Z_2)\\ &\le \frac{4}{n} + \frac{9}{m}\left(2 + \left(\frac{n}{m-1}\right) + \frac{n}{m}\right).\end{aligned} \tag{4.9}$$

For the second term in the right hand side of (4.6), set $\bar{V}_i := V_i - a$. By (4.7), $-a \le \bar{V}_i \le 1 - a$, and $|T_i| \le 1$. Hence by the case $k = 4$ of Lemma 3.4,

$$\text{Cov}(\bar{V}_1 T_2, \bar{V}_3 T_4) \le \mathbb{E}\,[\bar{V}_1 T_2 \bar{V}_3 T_4] \le \frac{3}{m}\left(2 + \left(\frac{n}{m-3}\right) + \frac{n}{m}\right).$$

By (4.8), we can always bound $\text{Cov}(\bar{V}_i T_j, \bar{V}_{i'} T_{j'})$ by 1. Hence, expanding $\text{Var}\sum_{(i,j):i\neq j} \bar{V}_i T_j$ in the same manner as with (6.25) below, yields

$$\text{Var}\sum_{(i,j):i\neq j} \frac{\bar{V}_i T_j}{n(n-1)} \le \frac{3}{m}\left(2 + \left(\frac{n}{m-3}\right) + \frac{n}{m}\right) + \frac{4}{n} + \frac{2}{n(n-1)}.$$

Using this with (4.6) and (4.9) yields

$$\text{Var}\left(\mathbb{E}\,[Y'' - Y|\mathscr{G}]\right) \le \frac{16}{n} + \frac{4}{n(n-1)} + \frac{24}{m}\left(2 + \frac{n}{m-3} + \frac{n}{m}\right).$$

This completes the proof of Proposition 4.1, and hence of Theorem 2.1. □

Proof of Theorem 2.2. Suppose n, m both go to infinity in a linked manner, in such a way that $n/m \to \alpha \in (0, \infty)$. Then it can be shown (see e.g. Theorem II.1.1 on pages 37-38 of [10]) that $\mathbb{E}\,Y \sim n(1 - e^{-\alpha})$, and

$$\text{Var}(Y) \sim n\left(e^{-\alpha}(1 - e^{-\alpha}) + e^{-2\alpha}(\alpha(1-\alpha))\right) = ng(\alpha)^2.$$

Substituting these asymptotic expressions into (2.2) and (2.3) and using the fact that in this asymptotic regime, $\left(n\eta(n,m)\right) \to 16 + 24\alpha(2 + 2\alpha)$, yields (2.5) and (2.6).

5 The non-uniform case: proof of Theorem 2.4

For this proof, we use the following notation. Given n, m, and the probability distribution $\mathbf{p}$ on $[m]$, let $X_1, X_2, \dots, X_{n+1}$ be independent $[m]$-valued random variables with common probability mass function $\mathbf{p}$. Given $i \le j \le n+1$, set $\mathbf{X}_j := (X_1, \dots, X_j)$ and

$$\mathbf{X}_{j\setminus i} := \begin{cases} (X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_j) & \text{if} \quad 1 < i < j \\ (X_2, \dots, X_j) & \text{if} \quad i = 1 \\ (X_1, \dots, X_{j-1}) & \text{if} \quad i = j. \end{cases}$$

Given any sequence $\mathbf{x} = (x_1, \dots, x_k)$, set

$$H(\mathbf{x}) = \sum_{i=1}^{k} \left(1 - \prod_{j \in [k] \setminus \{i\}} (1 - \mathbf{1}\{x_j = x_i\}) \right), \tag{5.1}$$

which is the number of non-isolated entries in the sequence $\mathbf{x}$, so that in particular, $Y = H(\mathbf{X}_n)$. We shall use the following consequence of Jensen's inequality: for all $k \in \mathbb{N}$,

$$(t_1 + t_2 + \cdots + t_k)^2 \le k(t_1^2 + \cdots + t_k^2), \quad \forall \ (t_1, \dots, t_k) \in \mathbb{R}^k. \tag{5.2}$$

We shall also use several times the fact that $-t^{-1}\ln(1-t)$ is increasing on $t \in (0,1)$ so that if (2.9) holds, then for all $x \in [m]$ we have

$$\ln(1 - p_x) \ge 11 p_x \ln(10/11) \ge -1.05 p_x \tag{5.3}$$

whereas $(1 - e^{-t})/t$ is decreasing on $t \in (0, \infty)$ so that by (2.7), for any $\alpha > 0$ and $x \in [m]$ we have

$$1 - e^{-\alpha n p_x} \ge (1 - e^{-\alpha\gamma})(n p_x / \gamma). \tag{5.4}$$

Proof of (2.13). We use Steele's variant of the Efron-Stein inequality [14]. This says, among other things, that when (as here) $X_1, \dots, X_{n+1}$ are independent and identically distributed random variables and H is a symmetric function on $\mathbb{R}^n$,

$$\begin{aligned}\mathrm{Var} H(X_1, \dots, X_n) &\le \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[(H(\mathbf{X}_n) - H(\mathbf{X}_{(n+1)\setminus i}))^2\right] \\ &= (n/2)\mathbb{E}\left[(H(\mathbf{X}_n) - H(\mathbf{X}_{(n+1)\setminus n}))^2\right].\end{aligned}$$

Hence, by the case $k = 2$ of (5.2),

$$\begin{aligned}\mathrm{Var} Y &\le n(\mathbb{E}\left[(H(\mathbf{X}_n) - H(\mathbf{X}_{n-1}))^2\right] + \mathbb{E}\left[(H(\mathbf{X}_{(n+1)\setminus n}) - H(\mathbf{X}_{n-1}))^2\right]) \\ &= 2n\mathbb{E}\left[(H(\mathbf{X}_n) - H(\mathbf{X}_{n-1}))^2\right].\end{aligned}$$

With M_j defined by (2.1), $H(\mathbf{X}_n) - H(\mathbf{X}_{n-1})$ is equal to $\mathbf{1}\{M_n \ge 1\} + \mathbf{1}\{M_n = 1\}$, so is nonnegative and bounded by $2\mathbf{1}\{M_n \ge 1\}$. Therefore,

$$\mathrm{Var}[Y] \le 8nP[M_n \ge 1] \le 8n\mathbb{E} M_n \le 8n^2 \sum_x p_x^2. \quad \square$$

Proof of (2.14). Construct a martingale as follows. Let $\mathscr{F}_0$ be the trivial σ-algebra, and for $i \in [n]$ let $\mathscr{F}_i := \sigma(X_1, \dots, X_i)$ and write $\mathbb{E}_i$ for conditional expectation given $\mathscr{F}_i$. Define martingale differences $\Delta_i = \mathbb{E}_{i+1} Y - \mathbb{E}_i Y$. Then $Y - \mathbb{E} Y = \sum_{i=0}^{n-1} \Delta_i$, and by orthogonality of martingale differences,

$$\mathrm{Var}[Y] = \sum_{i=0}^{n-1} E[\Delta_i^2] = \sum_{i=0}^{n-1} \mathrm{Var}[\Delta_i]. \tag{5.5}$$

We look for lower bounds for $E[\Delta_i^2]$. Note that

$$\Delta_i = \mathbb{E}_{i+1}[W_i], \quad \text{where} \quad W_i := H(\mathbf{X}_n) - H(\mathbf{X}_{(n+1)\setminus(i+1)}). \tag{5.6}$$

Recall from (2.1) that for $i < n$, M_{i+1} denotes the number of balls in the sequence of n balls, other than ball $i+1$, in the same position as ball $i+1$. Similarly, define M_{n+1} and M_k^i (for $k \in [n+1]$) by

$$M_{n+1} := \sum_{j \in [n]} \mathbf{1}\{X_j = X_{n+1}\}; \qquad M_k^i := \sum_{j \in [i] \setminus \{k\}} \mathbf{1}\{X_j = X_k\}, \tag{5.7}$$

so that M_{n+1} is the number of balls, in the sequence of n balls, in the same location as ball $n+1$, while M_k^i is similar to M_k, but defined in terms of the first i balls, not the first n balls.

Set $h_0(k) := \mathbf{1}\{k \geq 1\} + \mathbf{1}\{k = 1\}$. Then $H(\mathbf{X}_n) - H(\mathbf{X}_{n \setminus (i+1)}) = h_0(M_{i+1})$, and if $X_{n+1} \neq X_{i+1}$ then $H(\mathbf{X}_{(n+1) \setminus (i+1)}) - H(\mathbf{X}_{n \setminus (i+1)}) = h_0(M_{n+1})$, so that $W_i = h_0(M_{i+1}) - h_0(M_{n+1})$ in this case. For taking $\mathbb{E}_{i+1}$-conditional expectations, it is convenient to approximate $h_0(M_{i+1})$ and $h_0(M_{n+1})$ by $h_0(M_{i+1}^i)$ and $h_0(M_{n+1}^i)$ respectively. To this end, define

$$Z_i := W_i - (h_0(M_{i+1}^i) - h_0(M_{n+1}^i)). \tag{5.8}$$

Since $h_0(M_{i+1}^i)$ is $\mathscr{F}_{i+1}$-measurable, taking conditional expectations in (5.8) yields

$$h_0(M_{i+1}^i) = \mathbb{E}_{i+1}[W_i] + \mathbb{E}_{i+1}[h_0(M_{n+1}^i)] - \mathbb{E}_{i+1}[Z_i]. \tag{5.9}$$

Set $\delta := (288\gamma e^{1.05\gamma})^{-1}$. We shall show that for i close to n, in the sense that $n - \delta n \leq i \leq n$, the variances of the terms on the right of (5.9), other than $\mathbb{E}_{i+1}[W_i]$, are small compared to the variance of the left hand side, essentially because $\mathbb{E}_{i+1}[h_0(M_{n+1}^i)]$ is more smoothed out than $h_0(M_{i+1}^i)$, while $P[Z_i \neq 0]$ is small when i is close to n. These estimates then yield a lower bound on the variance of $\mathbb{E}_{i+1}[W_i]$.

First consider the left hand side $h_0(M_{i+1}^i)$. This variable takes the value 0 when $M_{i+1}^i = 0$, and takes a value at least 1 when $M_{i+1}^i \geq 1$. Hence,

$$\mathrm{Var}[h_0(M_{i+1}^i)] \geq (1/2)\min(P[M_{i+1}^i = 0], P[M_{i+1}^i \geq 1]). \tag{5.10}$$

For $i \leq n$, by (5.3) and (2.7),

$$\begin{aligned} P[M_{i+1}^i = 0] &= \sum_x p_x (1-p_x)^i \geq \sum_x p_x (1-p_x)^n \\ &\geq \sum_x p_x e^{-1.05 n p_x} \geq \gamma^{-1} e^{-1.05\gamma} \sum_x n p_x^2. \end{aligned} \tag{5.11}$$

For $i \geq (1-\delta)n$ we have $i \geq n/2$, so by (5.4) and the fact that $\gamma \geq 1$ by (2.7),

$$\begin{aligned} P[M_{i+1}^i \geq 1] &= \sum_x p_x (1 - (1-p_x)^i) \geq \sum_x p_x (1 - e^{-np_x/2}) \\ &\geq \sum_x p_x (1 - e^{-\gamma/2}) n p_x / \gamma \geq (1 - e^{-1/2})\gamma^{-1} \sum_x n p_x^2. \end{aligned} \tag{5.12}$$

Since $\gamma \geq 1$, and $e^{-1.05} < 1 - e^{-0.5}$, the lower bound in (5.11) is always less than that in (5.12), so combining these two estimates and using (5.10) yields

$$\mathrm{Var}[h_0(M_{i+1}^i)] \geq (1/2)\gamma^{-1} e^{-1.05\gamma} \sum_x n p_x^2, \qquad i \in [n - \delta n, n]. \tag{5.13}$$

Now consider the second term $\mathbb{E}_{i+1}[h_0(M^i_{n+1})]$ in the right hand side of (5.9). Set $N^i_x := \sum_{j=1}^i \mathbf{1}\{X_j = x\}$, and for $1 \le \ell \le i$ set M^i_ℓ to be $N^i_{X_\ell} - 1$. Also set $\tilde{h}_0(k) = (k+1)^{-1} h_0(k+1)$. Then, since $h_0(0) = 0$, we have that

$$
\begin{aligned}
\operatorname{Var} \mathbb{E}_{i+1}[h_0(M^i_{n+1})] &= \operatorname{Var} \sum_x p_x h_0(N^i_x) = \operatorname{Var} \sum_{j=1}^i p_{X_j} \tilde{h}_0(M^i_j) \\
&= \frac{i}{n^2} \operatorname{Var}[n p_{X_1} \tilde{h}_0(M^i_1)] + \frac{i(i-1)}{n^2} \operatorname{Cov}\left[n p_{X_1} \tilde{h}_0(M^i_1), n p_{X_2} \tilde{h}_0(M^i_2)\right]. \qquad (5.14)
\end{aligned}
$$

Suppose $i \le n$. Since $0 \le n p_{X_1} \tilde{h}_0(M^i_1) \le 2 n p_{X_1}$, (2.7) yields

$$
\frac{i}{n^2} \operatorname{Var}[n p_{X_1} \tilde{h}_0(M^i_1)] \le n^{-1} \mathbb{E}[4 n^2 p^2_{X_1}] = 4n \sum_x p_x^3 \le 4\gamma \sum_x p_x^2,
$$

while by Lemma 3.5 and (2.7),

$$
\operatorname{Cov}\left[n p_{X_1} \tilde{h}_0(M^i_1), n p_{X_2} \tilde{h}_0(M^i_2)\right] \le (3 + 3\gamma) 4\gamma^2 \sum_x p_x^2.
$$

Combining the last two estimates on (5.14) and using assumption (2.10) yields

$$
\begin{aligned}
\operatorname{Var}(\mathbb{E}_{i+1}[h_0(M^i_{n+1})]) &\le (1 + 3\gamma + 3\gamma^2) 4\gamma \sum_x p_x^2 \\
&\le (18\gamma e^{1.05\gamma})^{-1} n \sum_x p_x^2. \qquad (5.15)
\end{aligned}
$$

We turn to the third term in the right hand side of (5.9). As discussed just before (5.8), when $X_{n+1} \ne X_{i+1}$ we have $W_i = h_0(M_{i+1}) - h_0(M_{n+1})$, and it is clear from the definitions (5.6) and (5.7) that if $X_{n+1} = X_{i+1}$ then both W_i and $h_0(M^i_{i+1}) - h_0(M^i_{n+1})$ are zero, and therefore by (5.8),

$$
Z_i = (h_0(M_{i+1}) - h_0(M^i_{i+1}) - h_0(M_{n+1}) + h_0(M^i_{n+1})) \mathbf{1}\{X_{n+1} \ne X_{i+1}\}.
$$

By the conditional Jensen inequality,

$$
\operatorname{Var}(\mathbb{E}_{i+1}[Z_i]) \le \mathbb{E}[(\mathbb{E}_{i+1}[Z_i])^2] \le \mathbb{E}[Z_i^2].
$$

The random variable $h_0(M_{n+1}) - h_0(M^i_{n+1})$ lies in the range $[-2, 2]$ and is zero unless $X_j = X_{n+1}$ for some $j \in (i, n]$. Similarly, $h_0(M_{i+1}) - h_0(M^i_{i+1})$ lies in $[-2, 2]$ and is zero unless $X_j = X_{i+1}$ for some $j \in (i+1, n]$. Hence, using (5.2) and the definition of δ yields for $i \in [n - \delta n, n]$ that

$$
\begin{aligned}
\operatorname{Var}(\mathbb{E}_{i+1}[Z_i]) &\le 2(4P[M_{n+1} \ne M^i_{n+1}] + 4P[M_{i+1} \ne M^i_{i+1}]) \\
&\le 16\delta n \sum_x p_x^2 \le (18\gamma e^{1.05\gamma})^{-1} \sum_x n p_x^2. \qquad (5.16)
\end{aligned}
$$

By (5.9) and the case $k = 3$ of (5.2),

$$
\operatorname{Var}[h_0(M^i_{i+1})] \le 3(\operatorname{Var}(\mathbb{E}_{i+1}[W_i]) + \operatorname{Var}(\mathbb{E}_{i+1}[h_0(M^i_{n+1})]) + \operatorname{Var}(\mathbb{E}_{i+1}[Z_i])).
$$

Rearranging this and using (5.13), (5.15), and (5.16) yields the lower bound

$$\mathrm{Var}(\mathbb{E}_{i+1}[W_i]) \geq \left(\frac{1}{6} - \frac{2}{18}\right) \frac{e^{-1.05\gamma}}{\gamma} \sum_x np_x^2 = \frac{e^{-1.05\gamma}}{18\gamma} \sum_x np_x^2,$$

for $i \in [n - \delta n, n]$. Since the definition of δ, the condition (2.10) on n and the assumption $\gamma \geq 1$ guarantee that $n\delta \geq 2$, and since $\lfloor t \rfloor \geq 2t/3$ for $t \geq 2$, by (5.5) and (5.6) we have

$$\mathrm{Var}[Y] \geq \lfloor \delta n \rfloor (18\gamma e^{1.05\gamma})^{-1} n \sum_x p_x^2 \geq (\delta n)(27\gamma e^{1.05\gamma})^{-1} n \sum_x p_x^2$$

which is (2.14). □

6 Proof of Theorem 2.3

Proof of (2.8). Write σ for σ_Y, and for $t \in \mathbb{R}$ set $F(t) := P[(Y - \mathbb{E}Y)/\sigma \leq t]$. Set $z_0 := \sigma^{-1}(\lfloor \mathbb{E}Y \rfloor - \mathbb{E}Y)$, and set $z_1 := z_0 + (1-\varepsilon)/\sigma$, for some $\varepsilon \in (0,1)$. Then since Y is integer-valued, $F(z_1) = F(z_0)$. On the other hand, by the unimodality of the normal density,

$$\Phi(z_1) - \Phi(z_0) \geq (1 - \varepsilon)\sigma^{-1}(2\pi)^{-1/2} \exp(-1/(2\sigma^2))$$

so that D_Y is at least half the expression above. Making $\varepsilon \downarrow 0$ and using the fact that $e^{-1/(2\sigma^2)} \geq e^{-1/2}$ for $\sigma \geq 1$, gives us (2.8) in the case where $\sigma \geq 1$.

When $\sigma < 1$, we can take $z_2 \leq 0 \leq z_3$, with $z_3 = z_2 + 1$ and $F(z_3) = F(z_2)$. By the $68 - 95 - 99.7$ rule for the normal distribution, $\Phi(z_3) - \Phi(z_2) \geq 1/3$, so $D_Y \geq 1/6$, giving us (2.8) in the case where $\sigma < 1$. □

So in Theorem 2.3, the difficulty lies entirely in proving the upper bound in (2.11), under assumptions (2.9) and (2.10) which we assume to be in force throughout the sequel. By (2.10) we always have $n \geq 1661$.

As before, set $h_0(k) := \mathbf{1}\{k \geq 1\} + \mathbf{1}\{k = 1\}$. Define for nonnegative integer k the functions

$$h_1(k) := 1 - h_0(k) = \mathbf{1}\{k = 0\} - \mathbf{1}\{k = 1\};$$
$$h_2(k) := 2\mathbf{1}\{k = 1\} - \mathbf{1}\{k = 2\}; \qquad h_3(k) := \mathbf{1}\{k = 1\}.$$

The function $h_1(k)$ may be interpreted as the increment in the number of non-isolated balls should a ball in an urn containing k other balls be removed from that urn with the removed ball then deemed to be non-isolated itself. If $q = 0$ then the ball removed becomes non-isolated so the increment is 1, while if $q = 1$ then the other ball in the urn becomes isolated so the increment is -1.

The function $h_2(k)$ is chosen so that $h_2(k) + 2h_1(k)$ (for $k \geq 1$) is the increment in the number of non-isolated balls if two balls should be removed from an urn containing $k-1$ other balls, with both removed balls deemed to be non-isolated. The interpretation of h_3 is given later.

We shall need some further functions h_i which we define here to avoid disrupting the argument later on. For $x \in [m]$ and $k \in \{0\} \cup [n-1]$, let $\pi_k(x)$ be given by the π_k of (3.4) when $\nu = n - 1$ and

$p = p_x$. With the convention $0 \cdot \pi_{-1}(x) := 0 \cdot h_i(-1) := 0$, define

$$
\begin{aligned}
h_4(k,x) &:= \frac{k\pi_{k-1}(x)}{n-1} + \frac{(n-k-1)\pi_k(x)}{n-1} - 1; \\
h_5(k,x) &:= \pi_k(x)/(n-1), \quad h_6(k) := kh_2(k); \\
h_7(k,x) &:= h_3(k) + h_4(k,x) - \frac{k(2+h_4(k,x))h_1(k-1)}{n} \\
&\qquad - \frac{kh_5(k,x)(k-1)h_2(k-1)}{n}.
\end{aligned}
$$

For $i = 0,1,2,3,6$ define $h_i(k,x) := h_i(k)$. For each i define

$$\tilde{h}_i(k,x) := h_i(k+1,x)/(k+1). \tag{6.1}$$

Sometimes we shall write $\tilde{h}_i(k)$ for $\tilde{h}_i(k,x)$ when $i \in \{0,1,2,3,6\}$. Define $\|h_i\| := \sup_{k,x} |h_i(k,x)|$ and $\|\tilde{h}_i\| := \sup_{k,x} |\tilde{h}_i(k,x)|$.

Now we estimate some of the h_i functions. Since $\pi_0(x) = 1$ we have $h_4(0,x) = h_7(0,x) = 0$ for all x, which we use later. Also, by Lemma 3.3,

$$-1 \le h_4(k,x) \le 0 \tag{6.2}$$

and $h_7(1,x) = 1 + h_4(1,x)(1-n^{-1}) - 2/n$ so that $-1/n \le h_7(1,x) \le 1$. Also, since (2.10) implies $n \ge 1661$,

$$h_7(2,x) = h_4(2,x)(1+2n^{-1}) - \frac{4\pi_2(x)}{n(n-1)} + \frac{4}{n} \in [-1, 4/n].$$

Also, $h_3(3) = h_1(2) = 0$ so that by (6.2),

$$h_7(3,x) = \frac{6\pi_3(x)}{n(n-1)} + h_4(3,x) \in [-1,1],$$

again since $n \ge 1661$. For $k \ge 4$, $h_7(k,x) = h_4(k,x) \in [-1,0]$. Thus,

$$\|h_7\| \le 1; \quad \|\tilde{h}_7\| \le 1; \tag{6.3}$$

$$\|\tilde{h}_0\| = 2; \quad \|h_5\| \le (n-1)^{-1}; \quad \|h_6\| = 2. \tag{6.4}$$

The strategy to prove Theorem 2.3 is similar to the one already used in the uniform case, but the construction of a random variable with the distribution of Y', where Y' is defined to have the Y size biased distribution, is more complicated. As in the earlier case, by Lemma 3.2, if I is uniform over $[n]$ then the distribution of the sum Y conditional on $M_I > 0$ is the distribution of Y'. However, in the non-uniform case the conditional information that $M_I > 0$ affects the distribution of X_I. Indeed, for each i, by Bayes' theorem

$$P[X_i = x | M_i > 0] = \frac{p_x(1-(1-p_x)^{n-1})}{\sum_y p_y(1-(1-p_y)^{n-1})} =: \hat{p}_x. \tag{6.5}$$

Therefore the conditional distribution of $(X_1, \ldots, X_n)$, given that $M_i > 0$, is obtained by sampling X_i with probability mass function $\hat{p}$ and then sampling $\{X_j, j \in [n] \backslash \{i\}\}$ independently with probability mass function p, conditional on at least one of them taking the same value as X_i. Equivalently, sample the value of X_i, then M_i according to the binomial $\mathrm{Bin}(n-1, p_{X_i})$ distribution conditioned to be at least one, then select a subset $\mathscr{J}$ of $[n] \setminus \{i\}$ uniformly at random from sets of size M_i, let the values of $X_j, j \in \mathscr{J}$ be equal to X_i, and let the values of $X_j, j \notin \mathscr{J}$ be independently sampled from the distribution with the probability mass function of X given that $X \neq X_i$.

Thus a random variable Y'', coupled to Y and with the same distribution as Y', can be obtained as follows. First sample $X_1, \ldots, X_n$ independently from the original distribution $\mathbf{p}$, and set $\mathbf{X} = (X_1, \ldots, X_n)$; then select I uniformly at random from $[n]$. Then sample a further random variable X_0 with the probability mass function $\hat{p}$. Next, change the value of X_I to that of X_0; next let N denote the number of other values $X_j, j \in [n] \setminus \{I\}$ which are equal to X_0, and let $\pi_k = \pi_k(X_0)$ be defined by (3.4) with $v = n-1$ and $p = p_{X_0}$. Next, sample a Bernoulli random variable $\mathscr{B}$ with parameter π_N, and if $\mathscr{B} = 1$ change the value of one of the $X_j, j \in [n] \setminus \{I\}$ ($j = J$, with J sampled uniformly at random from all possibilities) to X_0. Finally, having made these changes, define Y'' in the same manner as Y in the original sum (2.1) but in terms of the changed variables. Then Y'' has the same distribution as Y' by a similar argument to that given around (4.1) in the uniform case.

Having defined coupled variables Y, Y'' such that Y'' has the Y size biased distribution, we wish to use Lemma 3.1. To this end, we need to estimate the quantities denoted B and Δ in that lemma. The following lemma makes a start. Let $\mathscr{G}$ be the σ-algebra generated by the value of $\mathbf{X}$, and for $x \in [m]$ let $N_x := \sum_{i=1}^n \mathbf{1}\{X_i = x\}$ be the number of balls in urn x.

Lemma 6.1. *It is the case that*

$$|Y'' - Y| \leq 3, \quad a.s. \tag{6.6}$$

and

$$\begin{aligned}\mathbb{E}[Y'' - Y | \mathscr{G}] = 2 &+ \left(\sum_x \hat{p}_x h_5(N_x, x) n^{-1} \sum_{i=1}^n h_6(M_i) \right) + \left(\sum_x \hat{p}_x h_7(N_x, x) \right) \\ &- \left(\sum_x \hat{p}_x h_4(N_x, x) \right) \left(n^{-1} \sum_{i=1}^n h_0(M_i) \right) - \frac{2}{n} \sum_{i=1}^n h_0(M_i). \end{aligned} \tag{6.7}$$

Proof. We have

$$\mathbb{E}[Y'' - Y | \mathscr{G}] = \mathbb{E}[Y'' - Y | \mathbf{X}] = \sum_x \hat{p}_x \mathbb{E}_x[Y'' - Y | \mathbf{X}], \tag{6.8}$$

where $\mathbb{E}_x[\cdot | \mathbf{X}]$ is conditional expectation given the value of $\mathbf{X}$ and given also that $X_0 = x$. The formula for $\mathbb{E}_x[Y'' - Y | \mathbf{X}]$ will depend on x through the value of N_x and through the value of p_x.

We distinguish between the cases where I is selected with $X_I = x$ (Case I) and where I is selected with $X_I \neq x$ (Case II). If $N_x = k$, then in Case I the value of N on which is based the probability $\pi_N(x)$ of importing a further ball to x is $k-1$ whereas in Case II this value of N is k. The probability of Case I occurring is k/n.

The increment $Y'' - Y$ gets a contribution of $h_1(M_i)$ from the moving of Ball i to x in Case II, and gets a further contribution of $h_1(M_j) + h_2(M_i)\mathbf{1}\{X_i = X_j\}$ if X_j is also imported to x from a location distinct from x. Finally, if $N_x = k$ the increment gets a further contribution of $h_3(k)$ from the fact that if there is originally a single ball at x, then this ball will no longer be isolated after importing at least one of balls I and J to x (note that $\pi_0(x) = 1$ so we never end up with an isolated ball at x). Combining these contributions, we have (6.6), and also that for values of $\mathbf{X}, x$ with $N_x = k$,

$$\begin{aligned}\mathbb{E}_x[Y'' - Y|\mathbf{X}] = h_3(k) + \frac{k\pi_{k-1}(x)}{n(n-1)} \sum_{\{j:X_j \neq x\}} h_1(M_j) + n^{-1} \sum_{\{i:X_i \neq x\}} h_1(M_i) \\ + \frac{\pi_k(x)}{n(n-1)} \sum_{(i,j):i\neq j, X_i \neq x, X_j \neq x} (h_1(M_j) + h_2(M_i)\mathbf{1}\{X_i = X_j\})\end{aligned}$$

where in the right hand side, the first sum comes from Case I and the other two sums come from Case II. Hence, if $N_x = k$ then

$$\begin{aligned}\mathbb{E}_x[Y'' - Y|\mathbf{X}] &= h_3(k) + \left(\frac{\pi_k(x)}{n(n-1)} \sum_{\{i:X_i \neq x\}} M_i h_2(M_i) \right) \\ &\quad + \left(\frac{k\pi_{k-1}(x)}{n(n-1)} + \frac{1}{n} + \frac{(n-k-1)\pi_k(x)}{n(n-1)} \right) \sum_{\{j:X_j \neq x\}} h_1(M_j) \\ &= h_3(k) + \frac{2 + h_4(k,x)}{n} \left(\left(\sum_{i=1}^{n} h_1(M_i) \right) - k h_1(k-1) \right) \\ &\quad + \frac{h_5(k,x)}{n} \left(\left(\sum_{i=1}^{n} h_6(M_i) \right) - k h_6(k-1) \right) \\ &= 2 + h_7(k,x) - \left(\frac{h_4(k,x)}{n} \sum_{i=1}^{n} h_0(M_i) \right) - \left(\frac{2}{n} \sum_{i=1}^{n} h_0(M_i) \right) \\ &\quad + \frac{h_5(k,x)}{n} \sum_{i=1}^{n} h_6(M_i).\end{aligned}$$

Then by (6.8) we have (6.7). □

The next lemma is based on the observation that since $h_4(0,x) = h_7(0,x) = 0$, two of the sums of the form $\sum_x$ in (6.7) are over non-empty urns so can be expressed as sums over the balls, i.e. of the form $\sum_{i=1}^{n}$. We need further notation. Set $\xi_i := \hat{p}_{X_i} \tilde{h}_4(M_i, X_i)$, and $T_j := h_0(M_j)$. Set $b := \mathbb{E}[T_j]$ (this does not depend on j), and $\bar{T}_j := T_j - b$. Again write $\sum_{(i,j):i\neq j}$ for $\sum_{i=1}^{n} \sum_{j\in[n]\setminus\{i\}}$.

Lemma 6.2. *It is the case that*

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[Y'' - Y|Y]) \leq 12(n-1)^{-2} + 3n^{-2}\mathrm{Var}\left(\sum_{(i,j):i\neq j} \xi_i \bar{T}_j \right) \\ + 3\mathrm{Var} \sum_{i=1}^{n} ([\tilde{h}_7(M_i, X_i) - (1-n^{-1}) b \tilde{h}_4(M_i, X_i)] \hat{p}_{X_i} \\ - [n^{-1} h_0(M_i)(2 + \hat{p}_{X_i} \tilde{h}_4(M_i, X_i))]).\end{aligned} \tag{6.9}$$

Proof. As in Section 4, (4.2) holds here too. So it suffices to prove (6.9) with the left hand side replaced by $\mathrm{Var}(\mathbb{E}[Y'' - Y|\mathscr{G}])$. Set

$$\rho(\mathbf{X}) := 2 + \sum_x \hat{p}_x h_5(N_x, x) n^{-1} \sum_{i=1}^n h_6(M_i). \tag{6.10}$$

Using (6.1), we reformulate the sums in (6.7) as follows. Since $h_7(0, x) = 0$,

$$\sum_x \hat{p}_x h_7(N_x, x) = \sum_{i=1}^n \hat{p}_{X_i} \left(\frac{h_7(M_i + 1, X_i)}{(M_i + 1)} \right) = \sum_{i=1}^n \hat{p}_{X_i} \tilde{h}_7(M_i, X_i). \tag{6.11}$$

Similarly, $\sum_x \hat{p}_x h_4(N_x, x) = \sum_{i=1}^n \hat{p}_{X_i} \tilde{h}_4(M_i, X_i)$ so that

$$\begin{aligned}\left(\sum_x \hat{p}_x h_4(N_x, x) \right) n^{-1} \sum_{i=1}^n h_0(M_i) = &\left(n^{-1} \sum_{i=1}^n \hat{p}_{X_i} \tilde{h}_4(M_i, X_i) h_0(M_i) \right) \\ &+ n^{-1} \sum_{(i,j):i \neq j} \hat{p}_{X_i} \tilde{h}_4(M_i, X_i) h_0(M_j).\end{aligned} \tag{6.12}$$

Substituting (6.10), (6.11) and (6.12) into (6.7) gives

$$\begin{aligned}\mathbb{E}[Y'' - Y|\mathscr{G}] = &\left(\sum_{i=1}^n \hat{p}_{X_i} \tilde{h}_7(M_i, X_i) \right) - \left(\sum_{i=1}^n \frac{h_0(M_i)}{n} (2 + \hat{p}_{X_i} \tilde{h}_4(M_i, X_i)) \right) \\ &+ \rho(\mathbf{X}) - n^{-1} \sum_{(i,j):i \neq j} \hat{p}_{X_i} \tilde{h}_4(M_i, X_i) h_0(M_j).\end{aligned} \tag{6.13}$$

The last sum in (6.13) can be rewritten as follows:

$$\sum_{(i,j):i \neq j} \xi_i T_j = \sum_{(i,j):i \neq j} \left(b\xi_i + \xi_i (T_j - b) \right) = (n-1) \left(\sum_{i=1}^n b\xi_i \right) + \sum_{(i,j):i \neq j} \xi_i \bar{T}_j.$$

Substituting into (6.13) yields

$$\begin{aligned}\mathbb{E}[Y'' - Y|\mathscr{G}] = \rho(\mathbf{X}) - &\left(n^{-1} \sum_{(i,j):i \neq j} \xi_i \bar{T}_j \right) + \sum_{i=1}^n (\hat{p}_{X_i} \tilde{h}_7(M_i, X_i) \\ &- (1 - n^{-1}) b \hat{p}_{X_i} \tilde{h}_4(M_i, X_i) - [n^{-1} h_0(M_i)(2 + \hat{p}_{X_i} \tilde{h}_4(M_i, X_i))]).\end{aligned}$$

By (6.10) and (6.4), $|\rho(\mathbf{X}) - 2| \leq 2(n-1)^{-1}$, so that $\mathrm{Var}(\rho(\mathbf{X})) \leq 4(n-1)^{-2}$. By (5.2), we then have (6.9) as asserted. ◻

Now we estimate $\hat{p}_x$. By (5.3), $(1 - p_y)^{n-1} \geq e^{-1.05 n p_y}$ for $y \in [m]$, so

$$1 - (1 - p_y)^{n-1} \leq 1 - e^{-1.05 n p_y} \leq 1.05 n p_y, \tag{6.14}$$

and by (5.4), (2.7) and the assumption that $n \geq 1661$ by (2.10),

$$\begin{aligned}1 - (1 - p_y)^{n-1} \geq 1 - e^{-0.9 n p_y} \geq (1 - e^{-0.9\gamma}) n p_y / \gamma \geq (1 - e^{-0.9}) n p_y / \gamma \\ \geq (0.55) n p_y / \gamma.\end{aligned} \tag{6.15}$$

By (6.5), (6.14) and (6.15), for all $x \in [m]$ we have that

$$\hat{p}_x \le \frac{2\gamma p_x^2}{\sum_y p_y^2}. \tag{6.16}$$

By (2.7) and (2.13), we have further that

$$\|\hat{\mathbf{p}}\| := \sup_x(\hat{p}_x) \le (2\gamma)\sup_x \frac{n^2 p_x^2}{\sum_y n^2 p_y^2} \le \frac{2\gamma^3}{n^2 \sum_y p_y^2} \tag{6.17}$$

$$\le \frac{16\gamma^3}{\mathrm{Var}Y}. \tag{6.18}$$

Also, by (6.16), (2.7) and (2.13),

$$\mathbb{E}\,\hat{p}_{X_1} \le \frac{2\gamma \sum_x p_x^3}{\sum_y p_y^2} \le \frac{2\gamma^2}{n}; \tag{6.19}$$

$$\mathbb{E}\,\hat{p}_{X_1}^2 \le \frac{4\gamma^2 \sum_x p_x^5}{(\sum_y p_y^2)^2} \le \frac{32n^2\gamma^2 \sum_x p_x^5}{(\mathrm{Var}Y)\sum_y p_y^2} \le \frac{32\gamma^5}{n\mathrm{Var}Y}. \tag{6.20}$$

Lemma 6.3. *With $C(\gamma)$ given by (2.12), it is the case that*

$$\mathrm{Var}(\mathbb{E}\,[Y'' - Y|Y]) \le \frac{(C(\gamma))^2}{\mathrm{Var}Y}. \tag{6.21}$$

Proof. We shall use the fact that by (2.13) $\mathrm{Var}Y \le 8n\gamma$ so

$$n^{-1} \le 8\gamma(\mathrm{Var}Y)^{-1}. \tag{6.22}$$

We estimate in turn the two variances in the right hand side of (6.9). First consider the single sum. Let S_i denote the ith term in that sum, i.e. set

$$S_i := [\tilde{h}_7(M_i, X_i) - (1 - n^{-1})b\tilde{h}_4(M_i, X_i)]\hat{p}_{X_i}$$
$$-h_0(M_i)[2 + \hat{p}_{X_i}\tilde{h}_4(M_i, X_i)]n^{-1},$$

and set $\bar{S}_i := S_i - \mathbb{E}\,S_i$. By (6.2) and (6.3), along with the fact that $h_0(k) \in [0, 2]$ so $0 \le b \le 2$, the coefficient of $\hat{p}_{X_i}$, in the definition of S_i, lies in the range $[-1, 3]$, while the coefficient of n^{-1} lies in the range $[-4, 0]$. Hence, $|S_i + \frac{2}{n}| \le 3\hat{p}_{X_i} + \frac{2}{n}$. By (5.2), (6.20) and (6.22),

$$3n\mathrm{Var}\left[S_1\right] \le 3n\mathbb{E}\left[\left(S_1 + (2/n)\right)^2\right] \le 6n\left(9\mathbb{E}\,[\hat{p}_{X_1}^2] + 4n^{-2}\right)$$
$$\le \frac{1728\gamma^5 + 192\gamma}{\mathrm{Var}Y}. \tag{6.23}$$

Also, in the notation of Lemma 3.5, if we write $\bar{S}_i = \psi(X_i, M_i)$ we have $\mathrm{rng}(\psi) \le 4\|\hat{\mathbf{p}}\| + 4n^{-1}$, and also $\|\psi\| \le 4\|\hat{\mathbf{p}}\| + 4n^{-1}$. Hence, by (6.17) and (6.18), followed by (2.13) and then (2.7),

$$\max(\mathrm{rng}(\psi), \|\psi\|) \le \left(4 + \sqrt{\frac{\mathrm{Var}Y \sum_x p_x^2}{2\gamma^6}}\right)\sqrt{\frac{32\gamma^6}{n^2\mathrm{Var}Y \sum_y p_y^2}}$$
$$\le \left(4 + \frac{2n\sum_x p_x^2}{\gamma^3}\right)\sqrt{\frac{32\gamma^6}{n^2\mathrm{Var}Y \sum_y p_y^2}} \le \left(4 + \frac{2}{\gamma^2}\right)\sqrt{\frac{32\gamma^6}{n^2\mathrm{Var}Y \sum_y p_y^2}}.$$

By Lemma 3.5,

$$3n^2\mathrm{Cov}\left[S_1,S_2\right]=3n^2\mathbb{E}\left[\bar{S}_1\bar{S}_2\right]\le 9(1+\gamma)\left(32\gamma^6/\mathrm{Var}Y\right)\left(16+16\gamma^{-2}+4\gamma^{-4}\right)$$
$$\le\frac{4608(\gamma^7+\gamma^6+\gamma^5+\gamma^4)+1152(\gamma^3+\gamma^2)}{\mathrm{Var}Y}.$$

Combining this with (6.23) yields

$$\begin{aligned}3\mathrm{Var}\sum_{i=1}^n S_i &= 3n\mathrm{Var}[S_1]+3n(n-1)\mathrm{Cov}[S_1,S_2]\\ &\le \left(\frac{100}{\mathrm{Var}Y}\right)\left(47\gamma^7+47\gamma^6+64\gamma^5+47\gamma^4+12\gamma^3+12\gamma^2\right).\end{aligned}\tag{6.24}$$

Consider now the double sum in (6.9). Writing $(n)_k$ for $n!/(n-k)!$, we have that

$$\begin{aligned}\mathrm{Var}\sum_{(i,j):i\ne j}\xi_i\bar{T}_j &= (n)_4\mathrm{Cov}(\xi_1\bar{T}_2,\xi_3\bar{T}_4)+(n)_2\left(\mathrm{Var}(\xi_1\bar{T}_2)+\mathrm{Cov}(\xi_1\bar{T}_2,\xi_2\bar{T}_1)\right)\\ &\quad+(n)_3\left(\mathrm{Cov}(\xi_1\bar{T}_2,\xi_1\bar{T}_3)+\mathrm{Cov}(\xi_2\bar{T}_1,\xi_3\bar{T}_1)+2\mathrm{Cov}(\xi_1\bar{T}_2,\xi_3\bar{T}_1)\right).\end{aligned}\tag{6.25}$$

For the first term of the right hand side of (6.25), observe that

$$\mathrm{Cov}(\xi_1\bar{T}_2,\xi_3\bar{T}_4)=\mathbb{E}\left[\xi_1\bar{T}_2\xi_3\bar{T}_4\right]-\mathbb{E}\left[\xi_1\bar{T}_2\right]\mathbb{E}\left[\xi_3\bar{T}_4\right]\le\mathbb{E}\left[\xi_1\bar{T}_2\xi_3\bar{T}_4\right],$$

and that $0\ge\xi_i\ge-\hat{p}_{X_i}$ by (6.2), while $0\le T_j\le 2$. So by Lemma 3.5, (6.17) and (6.18),

$$\begin{aligned}3n^2\mathrm{Cov}(\xi_1\bar{T}_2,\xi_3\bar{T}_4)&\le 12n^2\|\hat{\mathbf{p}}\|^2(9+9\gamma)\sum_x p_x^2\le 108(1+\gamma)\left(\frac{32\gamma^6}{\mathrm{Var}Y}\right)\\ &=(3456\gamma^6+3456\gamma^7)/\mathrm{Var}Y.\end{aligned}\tag{6.26}$$

Now consider the last term in (6.25). By (6.20),

$$3n\mathrm{Cov}(\xi_1\bar{T}_2,\xi_1\bar{T}_3)\le 3n\mathbb{E}\left[\xi_1^2\bar{T}_2\bar{T}_3\right]\le 12n\mathbb{E}\,\hat{p}_{X_1}^2\le\frac{384\gamma^5}{\mathrm{Var}Y},\tag{6.27}$$

while by (6.19) and (6.22),

$$\begin{aligned}3n(\mathrm{Cov}(\xi_2\bar{T}_1,\xi_3\bar{T}_1)+2\mathrm{Cov}(\xi_1\bar{T}_2,\xi_3\bar{T}_1))&\le 3n\mathbb{E}\,\bar{T}_1^2\xi_2\xi_3+6n\mathbb{E}\,\xi_1\bar{T}_2\xi_3\bar{T}_1\\ &\le 36n\mathbb{E}\left[\hat{p}_{X_1}\hat{p}_{X_2}\right]\le 144\gamma^4n^{-1}\le\frac{1152\gamma^5}{\mathrm{Var}Y}.\end{aligned}\tag{6.28}$$

The middle term in the right side of (6.25) is smaller; since $\mathbb{E}\left[\hat{p}_{X_1}\hat{p}_{X_2}\right]\le\mathbb{E}\left[\hat{p}_{X_1}^2\right]$ by the Cauchy-Schwarz inequality, (6.20) gives

$$3(\mathrm{Var}(\xi_1\bar{T}_2)+\mathrm{Cov}(\xi_1\bar{T}_2,\bar{T}_1\xi_2))\le 24\mathbb{E}\left[\hat{p}_{X_1}^2\right]\le\frac{768\gamma^5}{n\mathrm{Var}Y},$$

and since $n\ge 1661$ by (2.10), combined with (6.25), (6.26), (6.27), and (6.28), this shows that

$$3n^{-2}\mathrm{Var}\sum_{(i,j):i\ne j}\xi_i\bar{T}_j\le\left(\frac{100}{\mathrm{Var}Y}\right)(35\gamma^7+35\gamma^6+(15.9)\gamma^5).\tag{6.29}$$

Also, by (6.22) we obtain

$$12(n-1)^2 \le \left(\frac{12}{1660}\right)\left(\frac{n}{n-1}\right)n^{-1} \le \left(\frac{96\times 1661}{1660^2}\right)\frac{\gamma}{\mathrm{Var}Y} \le \frac{\gamma}{10\mathrm{Var}Y}.$$

Combining this with (6.9), (6.24) and (6.29) yields

$$\mathrm{Var}(\mathbb{E}\,[Y''-Y|Y]) \le \left(\frac{100}{\mathrm{Var}Y}\right)(82(\gamma^7+\gamma^6)+80\gamma^5+47\gamma^4+12(\gamma^3+\gamma^2)). \quad \square$$

Proof of Theorem 2.3. It remains to prove (2.11). By (6.14),

$$\mathbb{E}\,Y = n\sum_x p_x(1-(1-p_x)^{n-1}) \le 1.05n^2\sum_x p_x^2,$$

so that by (2.14),

$$\mu_Y/\sigma_Y^2 \le 8165\gamma^2 e^{2.1\gamma}. \tag{6.30}$$

We can apply (3.1) from Lemma 3.1 with $B=3$ by (6.6). By (6.30) and Lemma 6.3, this gives

$$D_Y \le \left(\frac{8165\gamma^2 e^{2.1\gamma}}{5}\right)\left(\sqrt{\frac{99}{\sigma_Y}+\frac{5C(\gamma)}{\sigma_Y}}+\frac{6}{\sqrt{\sigma_Y}}\right)^2,$$

which yields (2.11). $\square$

Acknowledgement. It is a pleasure to thank the Institut für Stochastik, Universität Karlsruhe, for their hospitality during the preparation of this manuscript, and Norbert Henze for valuable comments on an earlier draft.

References

[1] Barbour, A. D. and Gnedin, A. V. (2009) Small counts in the infinite occupancy scheme. *Electron. J. Probab.* **14**, 365-384. MR2480545

[2] Barbour, A. D., Holst, L. and Janson, S. (1992) *Poisson Approximation.* Oxford University Press, New York. MR1163825

[3] Chatterjee, S. (2008). A new method of normal approximation. *Ann. Probab.* **36**, 1584–1610. MR2435859

[4] Englund, G. (1981). A remainder term estimate for the normal approximation in classical occupancy. *Ann. Probab.* **9**, 684–692. MR0624696

[5] Feller, W. (1968) *An Introduction to Probability Theory and its Applications. Vol. I.* 3rd ed. John Wiley and Sons, New York. MR0228020

[6] Gnedin, A., Hansen, B. and Pitman, J. (2007) Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv.* **4**, 146–171. MR2318403

[7] Goldstein, L. and Penrose, M. D. (2008) Normal approximation for coverage models over binomial point processes. To appear in *Ann. Appl. Probab.* Arxiv: 0812.3084 v2.

[8] Hwang, H.-K. and Janson, S. (2008). Local limit theorems for finite and infinite urn models. *Ann. Probab.* **36**, 992–1022. MR2408581

[9] Johnson, N. L. and Kotz, S. (1977). *Urn Models and their Application: An approach to Modern Discrete Probability Theory.* John Wiley & Sons, New York. MR0488211

[10] Kolchin, V. F., Sevast'yanov, B.A. and Chistyakov, V. P. (1978). *Random Allocations.* Winston, Washington D.C. MR0471016

[11] Mikhailov, V. G. (1981). The central limit theorem for a scheme of independent allocation of particles by cells. (Russian) Number theory, mathematical analysis and their applications. *Trudy Mat. Inst. Steklov.* **157**, 138–152. MR0651763

[12] Quine, M. P. and Robinson, J. (1982). A Berry-Esseen bound for an occupancy problem. *Ann. Probab.* **10**, 663–671. MR0659536

[13] Quine, M. P. and Robinson, J. (1984) Normal approximations to sums of scores based on occupancy numbers. *Ann. Probab.* **12**, 794–804. MR0744234

[14] Steele, J. M. (1986) An Efron-Stein inequality for non-symmetric statistics. *Ann. Statist.* **14**, 753–758. MR0840528

[15] Vatutin, V. A. and Mikhailov, V. G. (1982) Limit theorems for the number of empty cells in an equiprobable scheme for the distribution of particles by groups. *Theory Probab. Appl.* **27**, 734–743. MR0681461