ASYMPTOTIC DISTRIBUTION FOR THE BIRTHDAY PROBLEM WITH MULTIPLE COINCIDENCES, VIA AN EMBEDDING OF THE COLLISION PROCESS

R. ARRATIA, S. GARIBALDI, AND J. KILIAN

Abstract. We study the random variable $B(c,n)$, which counts the number of balls that must be thrown into n equally-sized bins in order to obtain c collisions. The asymptotic expected value of $B(1,n)$ is the well-known $\sqrt{n\pi/2}$ appearing in the solution to the birthday problem; the limit distribution and asymptotic moments of $B(1,n)$ are also well known. We calculate the distribution and moments of $B(c,n)$ asymptotically as n goes to ∞ and $c = O(n)$.

We have two main tools: an embedding of the collision process — realizing the process as a deterministic function of the standard Poisson process — and a central limit result by Rényi.

Contents

1. Introduction 1
2. The classical occupancy process 3
3. The embedding 4
4. Uniform integrability in the uncentered case, $c = O(n)$ 11
5. Moments in the uncentered case, $c = o(n)$ 12
6. Results for collisions, based on duality 13
7. Uniform integrability and concentration in the centered case 16
8. Moments and variance in the centered case 20
References 21

1. Introduction

1.1. **Scientific motivation.** Imagine throwing balls into n equally-sized bins. One *collision* occurs whenever a ball lands in a bin that already contains a ball, so that a bin containing k balls contributes $k-1$ to the total number of collisions. This notion of collision is relevant for hash tables in computer science as in [Kn 3, §6.4] and for cryptology as in [KuS]; it is the definition of collision from [Kn 2, §3.3.2I, p. 69].

We study the random variable $B(c,n)$, which counts the number of balls you throw into n bins to produce c collisions. The classic birthday problem as described

2010 *Mathematics Subject Classification.* 60C05 (11Y16, 62E20).

Key words and phrases. birthday problem, collisions, Renyi, urn problem, size bias, chi distribution.

in [F, p. 33], [Mo, Problem 31], and [St] asks for the median of $B(1, 365)$. We define $B(0, n) = 0$. We have:

$$c + 1 \le B(c, n) \le c + n \quad \text{for } c = 1, 2, \ldots. \tag{1}$$

For $n = 1$, this forces $B(c, 1) = c + 1$, which makes sense, since with a single bin, the first ball does not make a collision but all subsequent balls do.

The variable $B(1, n)$ appears in the standard birthday problem for a year with n days, so it has already been well studied. Indeed,

$$B(1, n)/\sqrt{n} \Rightarrow L, \quad \text{and} \quad \mathbb{E}\, B(1, n) \sim \sqrt{n}\ \ \mathbb{E}\, L = \sqrt{\frac{\pi}{2}\, n}\,, \tag{2}$$

where the limit distribution of L is attributed to *L*ord Rayleigh, with $\mathbb{P}(L > t) = \exp(-t^2/2)$ for $t > 0$, and $\mathbb{E}\, L = \sqrt{\pi/2}$, see [D, Example 3.2.5]. On the other hand, for $c_n \to \infty$ with $c_n = o(n^{1/4})$ Kuhn and Struick [KuS, p. 221] show that

$$\mathbb{E}\, B(c_n, n) \sim \sqrt{2\, c_n\, n}\,,$$

which matches (2) apart from the coefficient of $c_n\, n$ inside the square root changing *from* $\pi/2$ *to* 2. Indeed, the impetus for this paper was the desire to explain how $\pi/2$ changes to 2, and the title of our initial writeup was "From $\pi/2$ to 2: $\pi/2, 9\pi/16, 75\pi/128, 1225\pi/2048, \ldots, 2$". See subsection 5.2 for more details.

1.2. **Quick survey of the contents.** We consider $B(c, n)$, the number of balls that must be thrown into n bins, in order to get a specified number c of collisions. To investigate this, we consider in Section 3 an embedding of the collision process into a standard Poisson process; the embedding may be of interest in its own right, and we give a variety of almost sure uniform error bounds, culminating in Theorem 8. Even the simplest process convergence, Theorem 3, which holds for all outcomes $\omega \in \Omega$, implies a process distributional limit, Corollary 4, which in turn gives the limit one-dimensional distributional limit: Corollary 5, which states that for fixed c, $B(c, n)/\sqrt{n} \Rightarrow \sqrt{2T_c}$, the chi distribution with $2c$ degrees of freedom. (That is, $2T_c$ is chi-square with $2c$ degrees of freedom. The chi distribution, although not as famous as the chi-squared distribution, appears naturally also in the tridiagonalization of random symmetric matrices, see [T, p. 79]. The chi distribution can be viewed as a generalized gamma distribution as in [JKB, pp. 388, 417].) The convergence result in Corollary 5, combined with a uniform integrability estimate in Section 4, gives the asymptotic mean and variance of $B(c, n)$ for fixed c, with details given in Section 5.2, in particular (37), (38), and (40).

We are mainly interested in the case where $c_n = o(n)$, because that is the case relevant for applications as in [KuS]. However, in analyzing the variance of $B(c, n)$, for $c \approx n^a$ with $1/2 \le a < 1$, our embedding is not an appropriate tool, and we were forced to work with duality and Rényi's central limit theorem for the number of empty boxes. This duality also *easily* handles the "central region", corresponding to $c_n/n \to \alpha_0 \in (0, \infty)$, hence we include such results in sections 6–8, such as Theorem 14 and Corollary 18. Note that the results in the last three sections concern a centered distribution $B(c, n) - \beta(c, n)$. Our penultimate result, Corollary 18, determines the moments of the centered distribution and the variance of $B(c, n)$ over a large range of choices for c. This, combined with the results of Section 5.2 extend the result from [KuS] to a much larger regime.

Our main results are new, despite the substantial existing literature on other occupancy problems, such as [JK], [KoSC], [H 86], [H 95], [GnHP], etc., and other

work on B such as [CaP]. (Although Theorem 8 could be recovered over a smaller regime by using Poisson approximation as in [ArGG] or [BHJ], cf. Remark 6 below.)

regime	convergence	uniform integrability	moments
fixed c	Theorem 5	Lemma 10	Corollaries 11, 13
$c = O(n^\alpha)$, $\alpha < 1$	Theorem 8	Lemma 10	Corollary 12
$c \to \infty$, $c/n \to \alpha_0 \in [0, \infty)$	Theorem 14	Lemma 17	Corollaries 12, 18, 19

TABLE 1. Summary of results concerning $B(c, n)$ as $n \to \infty$. The first two lines deal with the uncentered $B(c, n)$ and the last line with a centered version, $B(c, n) - \beta(c, n)$.

2. The classical occupancy process

The classical occupancy *problem* is specified in terms of a fixed number of balls, and a fixed number of equally likely bins. We choose the notation b balls and n bins, although the notation n balls and N bins, used for example by Rényi, is tempting, as it corresponds to the tradition, in statistics, of a sample of size n taken from a population of size N. The classical occupancy problem starts with independent and identically distributed $X_1, X_2, \ldots$, with $\mathbb{P}(X_t = i) = 1/n$ for $i = 1$ to n and $t \geq 1$, so that all n^b possible values for $(X_1, \ldots, X_b)$ are equally likely, and considers the distributions of $N_0 = N_0(b, n)$, the number of empty bins; $I = I(b, n)$, the number of occupied bins; and more generally, for each $k = 0, 1, 2, \ldots$, the distribution of $N_k = N_k(b, n)$, the number of bins with exactly k balls. Even at the level of describing the distribution of an individual $N_k(b, n)$, there is much to be said, see for example [KoSC, R, W, E, Mi, BG, BGI].

As a summary of the notation:

$$N_k(b, n) = \sum_{i=1}^{n} 1\left(k = \sum_{t=1}^{b} 1(X_t = i)\right) \tag{3}$$

is the number of bins containing exactly k balls, when b balls have been tossed into n bins. As a check:

$$\sum_{k \geq 0} N_k(b, n) = n \quad \text{and} \quad \sum_{k \geq 0} k\, N_k(b, n) = b.$$

The number of *occupied* bins, when b balls have been tossed into n bins, is

$$I(b, n) := n - N_0(b, n) = \sum_{k \geq 1} N_k(b, n), \tag{4}$$

and the number of collisions obtained is

$$C(b, n) := b - I(b, n) = \sum_{k \geq 1} (k - 1) N_k(b, n). \tag{5}$$

The classic occupancy *process* goes a little further: the number n of bins is fixed, and balls are tossed in succession, so that the count of occupied bins, $I(b, n)$, is determined by the locations $X_1, X_2, \ldots, X_b$ of the first b balls, and the entire process $(I(0, n), I(1, n), I(2, n), \ldots, I(b, n), \ldots)$ is determined by the locations $X_1, X_2, \ldots$ of the balls in $\{1, 2, \ldots, n\}$. Thanks to equally likely bins, the process

$(I(0,n), I(1,n), I(2,n), \ldots, I(b,n), \ldots)$ also has the structure of a birth process, with

$$\begin{aligned} \mathbb{P}(I(t+1,n) = i \mid I(t,n) = i) &= \frac{i}{n} \quad \text{and} \\ \mathbb{P}(I(t+1,n) = i+1 \mid I(t,n) = i) &= \frac{n-i}{n}. \end{aligned}$$

This idea, exploited by Rényi [R], may be considered as the foundation of our embedding, given in Section 3.

The collisions process can also be viewed as a birth process, with, for $b \geq c$,

$$\begin{aligned} \mathbb{P}(C(b+1,n) = c+1 \mid C(b,n) = c) &= \frac{b-c}{n} \quad \text{and} \\ \mathbb{P}(C(b+1,n) = c \mid C(b,n) = c) &= \frac{n-(b-c)}{n}. \end{aligned}$$

Finally, given the collision counting process, $(C(0,n), C(1,n), C(2,n), \ldots)$, the number of balls needed to get c collisions is defined by *duality*: for $c = 0, 1, 2, \ldots$,

$$B(c,n) := \inf\{b : C(b,n) \geq c\}, \tag{6}$$

where, of course, the infimum of the empty set is taken to be ∞.

3. The embedding

3.1. **Motivation and informal description.** Let Y denote the standard, rate 1 Poisson process. It has the property that $Y(t)$ is a Poisson random variable with expectation t. Define T_c to be the time of the c-th arrival in Y.

Let $f(p)$ be the amount of time one must run the process Y so that, with probability p, there is at least one arrival; by standard Poisson process calculations, for $0 \leq p < 1$, $1 - p = e^{-f(p)}$. We extend this to $f : [0,1] \to [0,\infty]$ given by

$$f(p) := -\log(1-p) = p + p^2/2 + p^3/3 + \cdots \quad \text{for } 0 \leq p < 1$$

and $f(1) := \infty$. Clearly f is strictly increasing, and maps its domain onto its range.

For fixed positive integer n, we now define a coupling of the random variables $B(c,n)$ and T_c for nonnegative integer c. To sample $B(1,n), B(2,n), \ldots, B(c_{\max}, n)$, we define state variables i, t and c, all of which are initially set to 0. For intuition, i denotes the number of occupied bins, t denotes the amount of time Y has run for, and c denotes the number of collisions. Our sampling algorithm repeats the following sequence of steps until $c = c_{\max}$.

(a) Run Y for *up to* $f(i/n)$ units of time, stopping immediately if there is an arrival; let a be the first arrival time if there is one.
(b) If there is no arrival, add $f(i/n)$ to t and increment i by 1; we call this a "miss" step; it corresponds to a ball being thrown without causing a collision. If there is an arrival, increment c, add a to t and put $(B(c,n), T_c) = (c+i, t)$; we call this a "hit" step.
(c) Return to step (a).

Note that conditioned on the arrivals $(T_1, T_2, ...)$, the sampling process described above is deterministic.

Also, the sequence $(B(1,n), B(2,n), \ldots)$ obtained from the sampling algorithm described above has the same distribution as the sequence of the same name defined in terms of throwing balls into bins. Indeed, if you throw a ball into 1 of n bins, i

of which are occupied, the probability of that throw causing a collision is i/n, the same as the probability of an arrival in step (a).

3.2. **Formal description.** To minimize notation, we will take the sample space to be the set of strictly increasing sequences of strictly positive real numbers, since such a sequence corresponds to the sequence of arrival times in the standard Poisson process:

$$\Omega = \{\omega = (t_1, t_2, \ldots) \in \mathbb{R}^\infty : 0 < t_1 < t_2 < \cdots\}. \tag{7}$$

Of course, in this setup, the random variable T_c is just the c-th coordinate, so for $\omega = (t_1, t_2, \ldots)$, $T_c(\omega) = t_c$.

Definition 1 (Formal specification of the embedding)**.** For any $n = 1, 2, \ldots$ and $\omega \in \Omega$, we *define* $B(c, n)(\omega)$ and $J(c, n)(\omega)$ for $c = 0, 1, 2, \ldots$ recursively, via

$$B(0, n) := 0, \quad J(0, n) := 0,$$

and for $c \geq 1$,

$$B(c, n)(\omega) := B(c-1, n)(\omega) + \inf\left\{ j : \left(\sum_{J(c-1,n)(\omega) \leq i < J(c-1,n)(\omega)+j} f(i/n) \right) \geq (T_c(\omega) - T_{c-1}(\omega)) \right\} \tag{8}$$

and

$$J(c, n)(\omega) := B(c, n)(\omega) - c.$$

(We think of $J(c, n) := B(c, n) - c$ as the number of *occupied* bins, when the number of balls tossed is just enough to have formed c collisions.)

We want to separate which properties of our coupling are *deterministic* from which are *distributional*. Hence to serve as the *range* for the coupling, with the notation $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$ for the nonnegative integers, we define

$$\mathcal{B} = \left\{ (b_0, b_1, b_2, \ldots) \in \mathbb{Z}_+^{\mathbb{Z}_+} : 0 = b_0 < b_1 < b_2 < \cdots \right\}. \tag{9}$$

Theorem 2. *With Ω given by (7) and $\mathcal{B}$ given by (9), for every value $n = 1, 2, \ldots$, the recursion in Definition 1 defines a map $\mathcal{C}_n$ with domain Ω and range $\mathcal{B}$:*

$$\begin{aligned} \mathcal{C}_n : \Omega &\to \mathcal{B} \\ \omega &\mapsto (B(0, n)(\omega), B(1, n)(\omega), B(2, n)(\omega), \ldots). \end{aligned} \tag{10}$$

When Ω is extended to $(\Omega, \mathcal{F}, \mathbb{P})$ so that $\omega = (t_1, t_2, \ldots)$ is distributed as the sequence of arrival times in the standard Poisson process, the resulting sequence $(B(0, n), B(1, n), B(2, n), \ldots)$ is distributed as the sequence for the classical occupancy model, with $B(c, n)$ being the number of balls that are needed to get c collisions, for $c = 0, 1, 2, \ldots$, as defined in section 2.

Proof. To see that $\mathcal{C}_n$ maps Ω *into* $\mathcal{B}$, we argue by induction on c. In (8), $\omega \in \Omega$ guarantees that $t_c - t_{c-1} \in (0, \infty)$, hence any j in the set on the right side of (8) is strictly positive. This set is nonempty, since $f(n/n) = \infty$, hence the infimum of the set is a positive integer at most n. Thus, for every $\omega \in \Omega$, for $c = 1, 2, \ldots$, $b_c - b_{c-1}$ is a positive integer, so $\mathcal{C}_n(\omega) \in \mathcal{B}$.

The distributional claim is proved as follows. In the classical occupancy model, we defined $C(b, n)$, the number of collisions after b balls have been tossed into

n bins, via (5). In terms of the map $\mathcal{C}_n$, here we *define* $C(b,n)$ by duality: $C(b,n) := \max\{c : B(c,n) \le b\}$. Note that $J(c,n)$ from Definition 1 corresponds to $I(C(B(c,n),n),n)$, the number of occupied bins after tossing the ball that causes the c-th collision.

For $b = 0, 1, 2, \ldots$, define the statement $S(b)$ to be (The joint distribution of $(C(0,n), C(1,n), \ldots, C(b,n))$, resulting from the map $\mathcal{C}_n$ applied to $(\Omega, \mathcal{F}, \mathbb{P})$, is identical to the joint distribution it would have in the classical occupancy model with n bins.) In proving $S(b)$ for all b, the last sentence of Section 3.1 provides the justification for $S(b)$ implies $S(b+1)$. □

3.3. **Preliminary analysis of the coupling.** We have

$$T_{c+1} = T_c + \sum_{i \in [I(B(c,n),n), I(B(c+1,n),n))} f(i/n) \quad + R(c+1) \tag{11}$$

where R, the random time involved in the last hit step, is limited by the fraction of bins which were occupied just before the $(c+1)$-st collision:

$$0 < R(c+1) \le f\left(\frac{I(B(c+1,n)-1,n)}{n}\right).$$

Define an auxiliary function

$$a(i,n) := \sum_{0 \le j < i} f(j/n).$$

Accumulating the hit or miss steps until $B(c,n)$ balls have been tossed, with c hits — equivalently, unwinding the recursion in (11) — gives

$$T_c = a(B(c,n) - c, n) + \sum_{x=1}^{c} R(x).$$

Write $b = B(c,n)$ so that

$$i = I(B(c,n),n) = b - c$$

is the number of occupied bins when the c-th collision is observed, and use $f(i/n)$ as an upper bound on $R(1), \ldots, R(c)$. This yields, for all $c, n \ge 1$,

$$a(i,n) \ \le T_c \le \ a(i,n) + c\, f(i/n), \quad \text{where } b = B(c,n) \text{ and } i = b - c = I(b,n). \tag{12}$$

The contribution from the first-order term of $f(p)$ to $a(i,n)$ is

$$a_1(i,n) := \sum_{0 \le j < i} j/n = \frac{i(i-1)}{2n} \ge \frac{(i-1)^2}{2n},$$

and

$$\left| a_1(i,n) - \frac{i^2}{2n} \right| \le \frac{i}{2n}. \tag{13}$$

If $i < n/2$, then for $j < i$, $p := j/n \in [0, 1/2)$ has $f(p) \le p + p^2$. Hence

$$\text{for } i < n/2, \ f(i/n) \le 2i/n \tag{14}$$

and, for $i < n/2$,

$$0 \le a(i,n) - a_1(i/n) = \sum_{0 \le j < i} (f(j/n) - j/n) \le \sum_{0 \le j < i} (j/n)^2 \le i^3/(3n^2). \tag{15}$$

A combination of (12) with (13), (14), and (15) is

$$\text{if } i := B(c,n) - c < n/2, \text{ then } \left| \frac{(B(c,n)-c)^2}{2n} - T_c \right| \le \frac{i}{2n} + \frac{i^3}{3n^2} + \frac{2ci}{n}. \tag{16}$$

3.4. **Theorems giving distributional and sure convergence.** To elucidate the structure of the embedding, we note that the basic convergence, given by (17) below, holds for all ω; it requires only knowing, for each $c = 1, 2, \ldots$, that $T_c(\omega) \in (0, \infty)$. It is a standard pattern in probability theory, that distributional convergence for an infinite-dimensional process, as in (21), is equivalent to the convergence of the finite dimensional distributions, akin to the distributional version of (17). But it is remarkable that even *one*-dimensional convergence, as in (20), implies joint convergence (17) and (21), with a dependent process limit — the reason is that we are dealing with an embedding, and can argue that there are no exceptional values ω; alternately, we could have argued about a.s. convergence, and noted that, with a discrete time setup, the countable union of null sets is again a null set.

Theorem 3. *Under the coupling given by Theorem 2,*

$$\left(\frac{B(1,n)}{\sqrt{2n}}, \frac{B(2,n)}{\sqrt{2n}}, \ldots \right) \to (\sqrt{T_1}, \sqrt{T_2}, \ldots) \tag{17}$$

for all $\omega \in \Omega$, as $n \to \infty$.

Proof. To prove (17), we note first that the usual topology on $\mathbb{R}^{\mathbb{N}}$ is the compact-open topology, so convergence is equivalent to having, for each fixed c, convergence under the projection into $\mathbb{R}^c$ using the first c coordinates. Thus, we prove that for fixed c, for every $\omega \in \Omega$, as $n \to \infty$,

$$\left(\frac{B(1,n)}{\sqrt{2n}}, \frac{B(2,n)}{\sqrt{2n}}, \ldots, \frac{B(c,n)}{\sqrt{2n}} \right) \to (\sqrt{T_1}, \sqrt{T_2}, \ldots, \sqrt{T_c}). \tag{18}$$

Write $i = B(c,n) - c$; this is random, varying with ω. Using the first half of (12), $a_1(i,n) \le a(i,n) \le T_c$ so

$$(i-1)^2 \le 2nT_c.$$

Hence $i = B(c,n) - c = O(\sqrt{n})$ as $n \to \infty$, for every $\omega \in \Omega$ — the implicit constants in the big Oh depend on $T_c(\omega)$. For sufficiently large n (again, depending on ω), $i < n/2$, so the upper bound in (16) applies, and

$$\left| \frac{(B(c,n)-c)^2}{2n} - T_c \right| \le \frac{i}{2n} + \frac{i^3}{3n^2} + \frac{2ci}{n} = O(n^{-1/2}), \tag{19}$$

using $i = O(\sqrt{n})$ and $c = O(1)$ to get the final conclusion in (19). Since $T_c(\omega) > 0$, (19) implies $(B(c,n) - c) \sim \sqrt{2nT_c}$. Since c is fixed, this implies

$$\frac{B(c,n)}{\sqrt{2n}} \to \sqrt{T_c}. \tag{20}$$

Finally, (20) implies (18); we could have shortened the proof, since (20) also implies (17) directly, but as discussed before stating this theorem, we want to highlight the unusual nature of the implication: one-dimensional convergence implies convergence of the infinite-dimensional joint distributions. □

As an immediate corollary to the second statement of Theorem 2, combined with Theorem 3, we get process distributional convergence, as stated formally by Corollary 4.

Corollary 4. *In the classical occupancy problem, tossing balls into n equally likely bins, as specified in section 2, as $n \to \infty$,*

$$\left(\frac{B(1,n)}{\sqrt{2n}}, \frac{B(2,n)}{\sqrt{2n}}, \dots \right) \Rightarrow (\sqrt{T_1}, \sqrt{T_2}, \dots), \tag{21}$$

where T_c is the time of the c-th arrival in a standard Poisson process. □

Remark. The joint distributional limit (21) in Corollary 4 of course gives the limit distribution under arbitrary continuous functionals on $\mathbb{R}^{\mathbb{N}}$, and there are many natural examples where the scaling by $\sqrt{2n}$ can be removed; for example, as $n \to \infty$

$$\mathbb{P}(B(3,n) - B(2,n) > B(1,n)) \to \mathbb{P}(\sqrt{T_3} - \sqrt{T_2} > \sqrt{T_1})$$

and

$$\mathbb{P}(B(1,n)B(5,n) > B^2(3,n)) \to \mathbb{P}(\sqrt{T_1 T_5} > T_3).$$

A result even weaker than Corollary 4 answers the basic question for collisions: What is the approximate distribution of the number $B(c,n)$ of balls that need to be tossed, to get c collisions, when there are n equally likely bins?

Corollary 5. *In the classical occupancy problem, tossing balls into n equally likely bins, as specified in section 2, for each fixed $c = 1, 2, \dots$, as $n \to \infty$,*

$$\frac{B(c,n)}{\sqrt{n}} \Rightarrow \sqrt{2T_c}, \tag{22}$$

where T_c is the time of the c-th arrival in a standard Poisson process.

The distribution of $2T_c$ is identical to the distribution of the sum of the squares of $2c$ standard normal random variables, and is well known as the chi-squared distribution with $2c$ degrees of freedom, or the gamma distribution with shape parameter c and scale parameter 2. Hence, $\sqrt{2T_c}$ is has the distribution of a chi random variable with $2c$ degrees of freedom.

Remark 6. One could count collisions in an alternative way: the number of collisions is the sum, over $1 \le i < j \le b$, of the indicator that balls i and j land in the same bin, with the overall effect that a bin containing k balls contributes $\binom{k}{2}$ to the total number of collisions, and the expected number of collisions is $\binom{b}{2}/n$. This method of counting lends itself to Poisson approximation; see for example [ArGG, p. 408], and as long as $b = o(n^{2/3})$, the difference between the two methods of counting may be considered as an error term, leading to an alternate proof of Corollary 4.

3.5. **Theorem giving almost sure asymptotics.** The next theorem *should* be paraphrased as "Almost surely, if $c \to \infty$ with $c = o(n)$, then $B(c,n) \sim \sqrt{2cn}$." The slightly sloppy paraphrase, "If $c \to \infty$ with $c = o(n)$, then $B(c,n) \sim \sqrt{2cn}$ a.s." is a weaker statement, since different sequences $c_1, c_2, \dots$ might have different null sets, and it is not easy to name a countable collection of sequences which cover all the sequences having $c_n \to \infty$ and $c_n/n \to 0$.

Theorem 7. *Under the coupling given by Theorem 2, the event $G = \{\lim_{c\to\infty} T_c/c = 1\}$ has probability 1. For all $\omega \in G$, for any sequence $c_1, c_2, \dots$ of positive integers such that $c_n \to \infty$ and $c_n/n \to 0$, we have*

$$\frac{B(c_n,n)}{\sqrt{2c_n n}} \to 1. \tag{23}$$

Proof. $\mathbb{P}(G) = 1$ by the strong law of large numbers. Write c for c_n and $i = (B(c,n) - c)$. For $\omega \in G$, the relation

$$(i-1)^2 \le 2nT_c$$

with $T_c \sim c = o(n)$ implies that $i^2 = O(nc) = o(n^2)$, hence $i = o(n)$. Hence for sufficiently large n (depending on the choice of $\omega \in G$), (16) applies, giving:

$$\left| \frac{(B(c,n)-c)^2}{2n} - T_c \right| \le \frac{i}{2n} + \frac{i^3}{3n^2} + \frac{2ci}{n} = o(c). \tag{24}$$

The equality is justified term-by-term, where the argument for the middle term is that $i^3/n^2 = (i^2/(nc))\ (i/n)\ c = O(1)\ o(1)\ c = o(c)$. Using $\omega \in G$, (24) implies that $i^2/(2n) \sim c$, equivalently $i^2 \sim 2nc$, equivalently $B(c,n) - c \sim \sqrt{2nc}$. Finally, since $c = o(\sqrt{2nc})$, this implies that $B(c,n) \sim \sqrt{2nc}$. □

3.6. **Theorem giving almost sure uniform convergence.**

Theorem 8. *Under the coupling given by Theorem 2, the event $H = \{T_c \asymp c\}$, that the ratio T_c/c is bounded away from zero and infinity, has probability 1. For all $\omega \in H$, there is uniform convergence, as given by the following:*

$$\sup_{c=o(n^{1/3})} \left| \frac{B^2(c,n)}{2n} - T_c \right| \to 0; \tag{25}$$

$$\sup_{c=o(n^{1/2})} \left| \frac{B(c,n)}{\sqrt{2n}} - \sqrt{T_c} \right| \to 0; \tag{26}$$

for any $C < \infty$ and $\alpha \in (0, 1/3)$,

$$\sup_{c \le Cn^\alpha} \left| \frac{B^2(c,n)}{2n} - T_c \right| = O(n^{(3\alpha-1)/2}); \tag{27}$$

and for any $C < \infty$ and $\alpha \in (0,1)$,

$$\sup_{c \le Cn^\alpha} \left| \frac{B(c,n)}{\sqrt{2n}} - \sqrt{T_c} \right| = O(n^{\alpha-(1/2)}). \tag{28}$$

Proof. Observe that in (16), each ingredient in the upper bound, $i = B(c,n)(\omega) - c = J(c,n)(\omega)$, and c itself, is a nondecreasing function of c. So immediately, we also have the stronger uniform statement if $i := B(c,n) - c$ satisfies $i < n/2$, then

$$\sup_{1 \le c' \le c} \left| \frac{(B(c',n)-c')^2}{2n} - T_{c'} \right| \le \frac{i}{2n} + \frac{i^3}{3n^2} + \frac{2ci}{n}. \tag{29}$$

The exact meaning of (25) is: for any sequence $c_1, c_2, \ldots$, such that $c_n/n^{1/3} \to 0$,

$$\text{for all } \omega \in H, \quad \lim_{n\to\infty} \sup_{1 \le c' \le c_n} \left| \frac{B^2(c',n)}{2n} - T'_c \right| = 0. \tag{30}$$

We may assume that $c_n \to \infty$, for if $\sup c_n < \infty$, then (30) holds simply as a corollary of (20) in Theorem 3.

Write $c = c_n$ and $i = B(c,n) - c$. Using the first half of (12), $a_1(i,n) \le a(i,n) \le T_c$ so

$$(i-1)^2 \le 2nT_c,$$

hence $i = O(\sqrt{nc}) = o(n)$ as $n \to \infty$, for every $\omega \in H$. For sufficiently large n (depending on ω), $i < n/2$, so the upper bound in (29) applies, and

$$\left| \frac{(B(c',n)-c')^2}{2n} - T_{c'} \right| \leq \frac{i}{2n} + \frac{i^3}{3n^2} + \frac{2ci}{n} =: u = O(\sqrt{c^3/n}) = o(1), \tag{31}$$

for $1 \leq c' \leq c$, noting that each of i^3/n^2 and ci/n just satisfies the $O(\sqrt{c^3/n})$ relation.

Once n is large enough that the upper bound u in (31) is less than 1, and large enough so that $T_c \leq Uc$ for some constant U, we have, for all $1 \leq c' \leq c = c_n$,

$$\frac{(B(c',n)-c')^2}{2n} \leq T_c + 1 \leq Uc + 1 =: t,$$

which implies that $(B(c',n)-c')^2 \leq 2nt$, hence $B(c',n) \leq \sqrt{2nt} + c' \leq \sqrt{2nt} + c$, hence expanding $(B(c',n)-c')^2$ and using the triangle inequality,

$$\begin{aligned} \left| \frac{(B(c',n))^2}{2n} - T_{c'} \right| &\leq \frac{2c'B(c',n) + c'^2}{2n} + \left| \frac{(B(c',n)-c')^2}{2n} - T_{c'} \right| \\ &\leq \frac{2c(\sqrt{2nt}+c) + c^2}{2n} + u = O(\sqrt{c^3/n}) \end{aligned} \tag{32}$$

for $1 \leq c' \leq c$. This completes the proof of (25), and simultaneously proves (27). We note that (34) below will provide an alternate proof of (25) and (27), without making use of the uniformity in (29).

Next, we prove (26). Consider the random variables $L := \inf_{c \geq 1} T_c/c$, $U := \sup_{c \geq 1} T_c/c$, so that $\omega \in H$ is precisely that $0 < L(\omega) \leq U(\omega) < \infty$. For every c, writing $i = B(c,n) - c$, the first half of (12), that $a_1(i,n) \leq a(i,n) \leq T_c$, yields $(i-1)^2 \leq 2nT_c$, so $\omega \in H$ yields the further $(i-1)^2 \leq 2nUc < \infty$, hence there is a random $K(\omega) < \infty$ such that for all $n, c \geq 1$, $i := B(c,n) - c \leq K\sqrt{nc}$. Note that this "big Oh" conclusion is weaker than the asymptotic in (23), but stronger in that it requires no condition on the growth of c relative to n. Next, for any $\varepsilon > 0$, imposing the growth condition $c \leq n^{1-\varepsilon}$, the relation $i := B(c,n) - c \leq K\sqrt{nc}$ implies that there exists $n_0(\omega) < \infty$ such that for all $n > n_0$, for all $c \leq n^{1-\varepsilon}$, $i := B(c,n) - c \leq n/2$, enabling (16), which we further bound using $i \leq K\sqrt{nc}$:

$$\left| \frac{(B(c,n)-c)^2}{2n} - T_c \right| \leq \frac{i}{2n} + \frac{i^3}{3n^2} + \frac{2ci}{n} \leq 2K^3(\sqrt{c/n} + \sqrt{c^3/n}) \leq 4K^3\sqrt{c^3/n}. \tag{33}$$

Next, similar to (32), we expand $(B(c,n)-c)^2$ and apply the triangle inequality; further using $B(c,n) \leq c + K\sqrt{nc}$ hence $2cB(c,n) + c^2 \leq 3c^2 + 2Kc\sqrt{nc}$ yields: for $\omega \in H$, for $n \geq n_0(\omega)$, for all $c \leq n^{1-\varepsilon}$,

$$\left| \frac{(B(c,n))^2}{2n} - T_c \right| \leq 4K^3\sqrt{c^3/n} + \frac{3c^2 + 2Kc\sqrt{nc}}{2n} \leq \frac{3c^2}{2n} + K'\sqrt{c^3/n} = O(\sqrt{c^3/n}), \tag{34}$$

and we have made the random implied constant in the big Oh more or less explicit.

What is the effect of applying square root, to both $B^2(c,n)/(2n)$ and to T_c, in (34)? Suppose that $c = n^\alpha$ with $\alpha \in (0,1)$; the random T_c is of order $T_c \asymp c = n^\alpha$, and suppose the amount of perturbation, $d := B^2(c,n)/(2n) - T_c$, is of order $d \asymp \sqrt{c^3/n} = n^{(3\alpha-1)/2}$. This situation has $d/c = n^{(\alpha-1)/2}$, which is $o(1)$ provided $\alpha < 1$. Using $d = o(c)$, we have $\sqrt{c+d} = \sqrt{c}\sqrt{1+d/c} = \sqrt{c}(1 + d/(2c) + O(d^2/c^2))$, leading to $\sqrt{c+d} - \sqrt{c} \sim \sqrt{c}\, d/(2c) = d/(2\sqrt{c}) \asymp d/\sqrt{c} \asymp n^{(3\alpha-1)/2}/n^{\alpha/2} = n^{\alpha - \frac{1}{2}}$.

The combination of this calculation, with the uniform upper bound (34), together with $\omega \in H$, proves both (26) and (28). □

4. Uniform integrability in the uncentered case, $c = O(n)$

4.1. **Uniform integrability.** Lemma 9 gives the crucial estimate, and Proposition 10 establishes uniform integrability, as required in the proof of Corollary 11, to get limit moments from the distributional convergence proved in Corollary 5.

Lemma 9. *Fix $0 < K < \infty$. For $n = 1, 2, \ldots$, for all c with $1 \le c \le Kn$, for all $t > \max(8K, 44)$,*

$$\mathbb{P}(B(c,n)/\sqrt{2cn} > \sqrt{t}) \le ce^{-ct/8}.$$

Proof. The restriction $t \ge 44$ implies that $t/8 \ge \log(2t)+1$. The restriction $1 \le c \le Kn$ implies $2(c+1)^2/(nc) \le 2(2c)^2/(nc) = 8c/n \le 8K$, so combined with $t \ge 8K$ we have $2(c+1)^2/(nc) \le t$, so $(c+1)^2 \le ntc/2$.

Put b for the floor of $\sqrt{2nct}$, so:

$$\mathbb{P}(B(c,n)/\sqrt{2cn} > \sqrt{t}) = \mathbb{P}(B(c,n) > b).$$

Using (6), this equals $\mathbb{P}(C(b,n) \le c-1)$, where $C(b,n)$ is the random variable denoting the number of collisions obtained after throwing b balls. The event $C(b,n) = y$ entails partitioning the set of b balls into $b-y$ blocks, i.e., disjoint nonempty subsets. For $y = 0, 1, \ldots, c-1$, we have

$$\mathbb{P}(C(b,n) = y) \le \binom{b}{y} b^y \frac{(n)_{b-y}}{n^b}$$

where the binomial coefficient is for choosing which y of the b balls were *colliders* (i.e., landed in a non-empty bin), the b^y term overcounts which ball each of the colliders collided with, $(n)_{b-y}$ describes the assignment of the $b-y$ blocks to bins, and there are n^b possible throws.

We substitute $\binom{b}{y} \le b^y/y!$ and $b^2 \le nt^2$ to obtain:

$$\mathbb{P}(C(b,n) = y) \le \frac{t^{2y}}{y!} \frac{(n)_{b-y}}{n^{b-y}}.$$

The second fraction is bounded above by $\exp(-\binom{b-y}{2}/n)$. Note that by our hypothesis on t, $c+1 < t\sqrt{n/2}$, hence

$$b - y - 1 = (b+1) - (y+1) - 1 \ge t\sqrt{n} - c - 1 \ge t\sqrt{n}/2$$

and $\binom{b-y}{2} \ge nt^2/8$. Omitting the $1/y!$ term gives

$$\mathbb{P}(C(b,n) = y) \le t^{2y} \exp(-nt^2/8) \le \exp(-(t^2/8 - 2y\log(t))).$$

As $t^2/8 - 2y\log(t) \ge t^2/8 - 2c\log(t) \ge t^2/16$, it follows that $\mathbb{P}(C(b,n) < c) \le ce^{-t^2/16}$, proving the claim. □

Lemma 10. *Fix $K < \infty$. For $k = 1, 2, \ldots$, the family*

$$\left\{ \left(\frac{B(c,n)}{\sqrt{cn}} \right)^k : n \ge 1, c \le Kn \right\}$$

is uniformly integrable.

Proof. The exponentially decaying uniform upper bound on the upper tail, given by Lemma 9, implies that for each $k = 1, 2, \ldots$, $\sup_n \sup_{c \le Kn} \mathbb{E}(B(c,n)/\sqrt{cn})^{k+1} < \infty$. This uniform boundedness of the $(k+1)$-st moments implies uniform integrability of the family of k-th powers. □

5. Moments in the uncentered case, $c = o(n)$

5.1. Corollaries of convergence together with uniform integrability.

Corollary 11. *For each fixed c, and for $k = 1, 2, \ldots$*

$$\mathbb{E}[B(c,n)^k] \sim (2n)^{k/2} \frac{\Gamma(c + k/2)}{\Gamma(c)} \quad \text{as } n \to \infty. \tag{35}$$

Proof. Combine the one-dimensional distribution convergence in Corollary 5, the uniform integrability in Lemma 10, and the formula for the moments of a chi-distributed random variable from [JKB, p. 421]. □

Corollary 12. *For $c = c_n$ with $c_n \to \infty$ and $c_n/n \to 0$, and for $k = 1, 2, \ldots$,*

$$\mathbb{E}[B(c_n,n)^k] \sim (2nc_n)^{k/2} \quad \text{as } n \to \infty.$$

Proof. Combine the almost sure limit in Theorem 7 with the uniform integrability in Lemma 10. □

5.2. Explicit asymptotics for the mean and variance of $B(c,n)$, fixed c.

Define, for $c = 1, 2, \ldots$,

$$\gamma(c) := \frac{\mathbb{E}[\sqrt{T_c}]}{\sqrt{c}} = \frac{\Gamma(c + 1/2)}{\sqrt{c}\,\Gamma(c)}, \tag{36}$$

so that (35), specialized to $k = 1$, may be paraphrased as

$$\mathbb{E}\, B(c,n) \sim \gamma(c)\sqrt{2nc} \quad \text{for fixed } c = 1, 2, \ldots \tag{37}$$

It is an exercise using the material in [AAR, Chap. 1] to show that $\gamma(c)$ is increasing.

Calculating explicitly $\gamma(c)$ for a few values of c gives:

c	1	2	3	4	5	$\cdots$	∞
$\gamma(c)$	$\sqrt{\frac{\pi}{4}}$	$\sqrt{\frac{9\pi}{32}}$	$\sqrt{\frac{75\pi}{256}}$	$\sqrt{\frac{1225\pi}{4096}}$	$\sqrt{\frac{19845\pi}{65536}}$	$\cdots$	1

where the limit $\gamma(c) \to 1$ is the $k = 1$ case of the well-known formula

$$\lim_{c\to\infty} \frac{\Gamma\left(c + \frac{k}{2}\right)}{\Gamma(c)\,\sqrt{c}^{\,k}} = 1.$$

So $\gamma(c)$ increases from about 0.886 to 1. And c need not be very big for $\gamma(c)$ to be close to 1; for $c \ge 13$, $\gamma(c) > 0.99$.

Plugging numerators into Sloane's [Slo] gives a hit on sequence A161736, "Denominators of the column sums of the BG2 matrix", suggesting the formula

$$\gamma(c) = \frac{(2c)!}{2^{2c}(c!)^2}\sqrt{\pi c} \tag{38}$$

for integer c, which is easily verified using Legendre's duplication formula $\Gamma(c + 1/2) = \Gamma(2c)\sqrt{\pi}2^{1-2c}/\Gamma(c)$. Now Stirling's approximation can be applied to show how $\gamma(c)$ approaches 1 as c grows:

$$\gamma(c) = 1 - \frac{1}{8c} + \frac{1}{128c^2} + \frac{5}{1024c^3} - \frac{21}{32768c^4} + O(1/c^5). \tag{39}$$

For fixed c, Corollary 11 also gives the asymptotic variance.

Corollary 13. *For constant c,*

$$\lim_{n\to\infty} \frac{\operatorname{Var} B(c,n)}{n} = 2c(1-\gamma(c)^2). \tag{40}$$

Proof. The implication is immediate, by taking $k=1$ and $k=2$ in (35). □

Calculating explicitly the expression in (40) for the first few values of c gives

c	1	2	3	4	5	$\cdots$	∞
$\lim_{n\to\infty} \frac{\operatorname{Var} B(c,n)}{n}$	0.4292	0.4567	0.4777	0.4835	0.4869	$\cdots$	0.5

where the last column of the table means only that

$$\lim_{c\to\infty}\lim_{n\to\infty} \frac{\operatorname{Var} B(c,n)}{n} = \frac{1}{2}. \tag{41}$$

Indeed, the limit of $(\operatorname{Var} B(c,n))/n$ has the series expansion

$$\lim_{n\to\infty} \frac{\operatorname{Var} B(c,n)}{n} = \frac{1}{2} - \frac{1}{16c} - \frac{1}{64c^2} + \frac{5}{1024c^3} + \frac{23}{4096c^4} + \cdots.$$

6. Results for collisions, based on duality

In the remainder of the paper, we find the asymptotic variance of $B(c,n)$ when $c = c_n \to \infty$ with $c/n \to \alpha_0 \in [0,\infty)$. Our method is to combine duality with Rényi's central limit result for the number of empty bins, to get a normal limit for $B(c,n)$ (this section), and to prove a concentration result to get uniform integrability (section 7), so that the normal limit governs the asymptotic variance (see section 8).

6.1. **History: Weiss and Rényi.** Weiss in 1958, [W], proved a central limit theorem for $N_0(b,n)$ in the "central regime", where $b,n \to \infty$ with $b \asymp n$, i.e., with ratio bounded away from zero and infinity. Weiss explicitly stated that N_0 is asymptotically normal, and implicit in this, together with his proof, is that the interpretation of asymptotic normality involves subtracting off the mean of N_0 and dividing by the standard deviation of N_0, i.e., Weiss proved that

$$\frac{N_0 - \mathbb{E}\, N_0}{\sqrt{\operatorname{Var} N_0}} \Rightarrow Z.$$

Renyi in 1962, [R], gave 3 proofs and went a little further. He gave a nice explicit expression to approximate the mean and variance, and, in his Theorem 2, "the third proof," extended the result to the case $b = o(n), b^2/n \to \infty$.

For motivation, suppose that $b/n \to \lambda$, so that a fixed number λ serves as the limit average number of balls per bin, $e^{-\lambda}$ is the limit probability that a given bin is empty, and the number of empty bins is asymptotic to $ne^{-\lambda}$. With

$$d(x) := e^{-x}\left(1-(1+x)e^{-x}\right) \quad \text{and} \quad \sigma^2(b,n) := d(b/n), \tag{42}$$

Renyi observes that $\operatorname{Var} N_0(b,n) = n\,\sigma^2(b,n)(1+O(b/n^2))$ — so $n\,\sigma^2$ is not the variance, but rather, a nice approximation of the variance. Likewise, $ne^{-b/n}$ is not $\mathbb{E}\, N_0$, but rather, a nice approximation. We remark that for the case $x \to 0$,

$$d(x) = e^{-2x}(e^x - 1 - x) \sim x^2/2. \tag{43}$$

A restatement of Rényi's Theorem 2, using our notation, is: *If* $b, n \to \infty$ *with* $b = O(n)$ *and* $b^2/n \to \infty$, *then*

$$\frac{N_0(b,n) - ne^{-b/n}}{\sqrt{n}\,\sigma(b,n)} \Rightarrow Z. \tag{44}$$

More advanced versions of Renyi's theorem, with concrete error bounds, are given by [E, Mi, BG], but for our purpose, to get a central limit for the number of collisions, using duality as in the proof of Theorem 14, Rényi's (44) is ideal.

6.2. **Unified normal limit, using duality.** The treatment in this section is *unified* in the sense that it handles both the regime $c_n \to \infty$ with $c_n = o(n)$, where the number of balls per bin approaches zero, and the regime $c_n \sim \alpha_0\, n$ with $\alpha_0 \in (0, \infty)$, where the number of balls per bin approaches a limit $\lambda_0 \in (0, \infty)$.

We define the function $w\colon [0,\infty) \to [0,\infty)$ via

$$w(x) := e^{-x} + x - 1. \tag{45}$$

It is easily checked that w is strictly increasing and onto, with $w(0) = 0$. Set

$$w^{-1}(\delta) := 1 + \delta + W(-e^{-1-\delta}) = -\log(-W(-e^{-1-\delta})) \quad \text{for } \delta \geq 0,$$

where W denotes the principal real-valued Lambert function, i.e., the concave increasing solution of $W(z)e^{W(z)} = z$ mapping $[-e^{-1}, \infty)$ onto $[-1, \infty)$ described in [Cor$^+$]. We find $w(w^{-1}(\delta)) = \delta$, compare p. 332 of ibid.

Given $c, n > 0$, we define

$$\beta(c,n) := n w^{-1}(c/n). \tag{46}$$

Applying the series expansion of $W(x)$ as x decreases to $-e^{-1}$ given in [Cor$^+$, (4.22)] gives

$$w^{-1}(\delta) = \sqrt{2\delta} + \frac{\delta}{3} + \frac{\delta^{3/2}}{9\sqrt{2}} + \cdots \quad \text{for small } \delta \geq 0. \tag{47}$$

In particular, when c_n is a function of n so that $c_n/n \to 0$, we have

$$\beta(c_n, n) = \sqrt{2c_n n} + \frac{c_n}{3} + \frac{1}{9\sqrt{2}} \frac{c_n^{3/2}}{\sqrt{n}} + \cdots$$

and $\beta(c_n, n) \sim \sqrt{2c_n n}$.

We define a continuous function $g(x)$ on $[0, \infty)$ via

$$g(0) := \frac{1}{\sqrt{2}} \quad \text{and} \quad g(x) := \frac{\sqrt{d(x)}}{1 - e^{-x}} \text{ for } x > 0, \tag{48}$$

where $d(x)$ is as in (42).

Theorem 14. *Suppose* $\lim_{n\to\infty} c_n = \infty$ *and* $\lim_{n\to\infty} c_n/n = \alpha_0 \in [0, \infty)$. *Then, with* $\beta(c,n)$ *defined by* (46) *to give the centering, with* w *as defined by* (45), $\lambda_0 := w^{-1}(\alpha_0)$, *and* g *as defined by* (48) *to give the scaling, we have the following convergence in distribution to the standard normal random variable* Z*:*

$$\frac{B(c_n, n) - \beta(c_n, n)}{g(\lambda_0)\sqrt{n}} \Rightarrow Z. \tag{49}$$

Proof. Let $c = c_n$ be given, with $c \to \infty$ and $c/n \to \alpha_0 \in [0, \infty)$. Write $\beta = \beta(c, n), \lambda = \beta/n$ so that (46) says that $c/n = w(\beta/n)$, i.e., $c = nw(\lambda)$, i.e.,

$$c = n\, e^{-\lambda} + n\lambda - n. \tag{50}$$

For fixed real y, let

$$b = \beta + y\sqrt{n}, \text{ so } b/n = \lambda + y/\sqrt{n}. \tag{51}$$

In terms of the cumulative distribution function Φ for the standard normal, so that $\mathbb{P}(Z > y) = \mathbb{P}(Z < -y) =: \Phi(-y)$, (49) is equivalent to showing that for each fixed y, $\mathbb{P}(B(c, n) > b) \to \mathbb{P}(g(\lambda_0)\, Z > y) = \Phi(-y/g(\lambda_0))$.

To enable us to use Rényi's result (44), we need to check that $b > 0$ for sufficiently large n, that b/n is bounded, and that $b^2/n \to \infty$. As $b/n = \beta/n + y/\sqrt{n} = w^{-1}(c_n/n) + y/\sqrt{n}$ and c_n is $O(n)$, b/n is bounded. Further, $\lim_{n\to\infty} \sqrt{n} w^{-1}(c_n/n) = \infty$ — this is obvious if $\alpha_0 \neq 0$ and follows from $c_n \to \infty$ and (47) if $\alpha_0 = 0$ — hence $\lim_{n\to\infty} b/\sqrt{n} = \infty$.

Recall, from (4) and (5), that $C(b, n)$, the number of collisions obtained after throwing b balls, and $N_0(b, n)$ for the number of empty bins remaining after throwing b balls, are related by

$$C(b, n) = b - (n - N_0(b, n)).$$

So by duality we have:

$$\mathbb{P}(B(c, n) > b) = \mathbb{P}(C(b, n) < c) = \mathbb{P}(N_0(b, n) < n - (n\lambda + y\sqrt{n}) + c).$$

Applying (50), we find:

$$\mathbb{P}(B(c, n) > b) = \mathbb{P}(N_0(b, n) < ne^{-\lambda} - y\sqrt{n}) \tag{52}$$

$$= \mathbb{P}\left(N_0(b, n) - ne^{-b/n} < y\sqrt{n}(e^{-\lambda} - 1) + O(1)\right). \tag{53}$$

We remark that $\sqrt{n}\,\sigma(b, n) = \sqrt{n\, d(b/n)} \to \infty$. Indeed, as b/n is bounded and $d(x)$ is nonzero for positive x, it suffices to check this in the case where $b/n \to 0$, where it follows from (43). Therefore, dividing both sides of the inequality in (53) gives

$$\mathbb{P}(B(c, n) > b) = \mathbb{P}\left(\frac{N_0(b, n) - ne^{-b/n}}{\sqrt{n}\,\sigma(b, n)} < \frac{y(e^{-\lambda} - 1)}{\sigma(b, n)}\right).$$

By (44), to complete the proof of the theorem it remains only to verify that

$$\sigma(b, n)/(e^{-\lambda} - 1) \to -g(\lambda_0). \tag{54}$$

If $\alpha_0 = 0$ then $\lambda_0 = w^{-1}(0) = 0$. We have $\beta(c, n) \sim \sqrt{2cn}$ hence $c \to \infty$ implies $b \sim \beta$. Using $\lambda = \beta/n \to 0$ we get $\sigma^2(\beta, n) \sim \lambda^2/2$ by (43), so $\sigma(\beta, n) \sim \lambda/\sqrt{2} = \lambda g(\lambda_0)$. It follows similarly, since $b \sim \beta$, that $\sigma(b, n) \sim \sigma(\beta, n) \sim \lambda g(\lambda_0)$. And of course $e^{-\lambda} - 1 \sim -\lambda$; (54) follows.

If $\alpha_0 > 0$ then $\lambda \to \lambda_0 = w^{-1}(\alpha_0) > 0$, and $\sigma^2(b, n) \to d(\lambda_0)$, a number, verifying (54). This concludes the proof of the theorem. □

7. Uniform integrability and concentration in the centered case

7.1. **Overview: background on concentration inequalities.** In order to conclude, from Theorem 14, that the variance of $B(c,n)$ is asymptotic to $n\,g(\lambda_0)^2$, we need uniform integrability. As in the proof of Theorem 14, fluctuations for $B(c,n)$, the number of balls that must be tossed to get c collisions, are related via duality to fluctuations of

$$C(b,n) = b - (n - N_0(b,n)), \tag{55}$$

the number of collisions resulting from tossing b balls. Hence we investigate concentration bounds for $N_0(b,n)$, or directly equivalently, concentration bounds for $C(b,n)$.

The central region, with c and b both of order n, is relatively easy to handle. In contrast, it took much effort to understand the region of main interest here: $c \to \infty$ with $c = o(n)$, so that $c = o(b)$ and $b = o(n)$. The following three random variables with exactly the same variance, but in the region of main interest, their expectations have different order of growth. From large to small, they are:

$$\begin{aligned} N_0(b,n) \quad &\text{with} \quad \mathbb{E}\,N_0(b,n) \sim n, \\ n - N_0(b,n) \quad &\text{with} \quad \mathbb{E}(n - N_0(b,n)) \sim b, \text{ and} \\ C(b,n) = b - (n - N_0(b,n)) \quad &\text{with} \quad \mathbb{E}\,C(b,n) \sim c. \end{aligned}$$

Consider applying Azuma's inequality for martingales with bounded differences, with random variables $X_1, \ldots, X_b \in \{1, 2, \ldots, n\}$ to say the destination bin for each ball, and sigma-algebras $\mathcal{F}_i := \sigma(X_1, \ldots, X_i)$ to carry the information known when the first i balls have been tossed, and $M_i := \mathbb{E}(N_0(b,n)|\mathcal{F}_i)$ for the martingale. It is obvious that $|M_i - M_{i-1}| \le 1$ for $i = 1$ to b, so Azuma gives the bounds

$$\begin{aligned} \mathbb{P}(N_0(b,n) - \mathbb{E}\,N_0(b,n) \ge t) &\le \exp(-t^2/(2b)), \\ \mathbb{P}(N_0(b,n) - \mathbb{E}\,N_0(b,n) \le -t) &\le \exp(-t^2/(2b)). \end{aligned} \tag{56}$$

For the central region, where c and b are both of order n, these bounds give us enough concentration to prove the desired UI result. In contrast, for the region of main interest to us, with $c = o(n)$ and $b \sim \sqrt{2cn}$, the bounds (56) are inadequate. For the region of main interest, we use a bounded size bias coupling whose existence is provided in the next subsection.

7.2. **Bounded size biased couplings for the number of collisions.**

Proposition 15. *Consider the occupancy model, generalized so that the locations of the balls, $X_1, \ldots, X_b \in \{1, 2, \ldots, n\}$, are still mutually independent, but not necessarily uniformly distributed, nor even identically distributed. There is a coupling of $C(b,n)$ with its size biased version $C'(b,n)$ such that $C'(b,n) - C(b,n) \in \{0,1,2\}$ for all outcomes. Furthermore, if each X_i is uniformly distributed on the boxes $1, \ldots, n$, then there is a coupling such that $C'(b,n) - C(b,n) \in \{0,1\}$ for all outcomes.*

Proof. We will consistently use the following notation: $1 \le i < j \le b, 1 \le k \le n$, so that i and j refer to balls, and k to bins, and i is tossed before j. Note this entails $j \ge 2$.

Write $Z_{ik} \equiv Z_{i,k} = 1(X_i = k)$ for the indicator that ball i lands in box k. The indicator that ball j lands in box k *and* accounts for a new collision — because at

least one earlier ball had already landed in box k — is

$$Y_{jk} = Z_{jk}\, 1(Z_{1k} + Z_{2k} + \cdots + Z_{j-1,k} > 0)$$

and the indicator that ball j, when it lands, accounts for a new collision, is

$$W_j = \sum_{k=1}^{n} Y_{jk}.$$

Hence the total number of collisions, when b balls are tossed into n boxes, can be expressed as

$$C(b,n) = \sum_{j=2}^{b} W_j \tag{57}$$

or

$$C(b,n) = \sum_{j=2}^{b} \sum_{k=1}^{n} Y_{jk} \tag{58}$$

We recall some basics about size bias, as presented by [ArGK, equations (15) and (12)]. First, for sums such as (57) or (58), the size bias distribution is naturally expressed as a mixture, with weights proportional to the contribution that a single term makes to the expected sum, of the sum for the process where the joint distribution of summands is biased in the direction of the chosen summand. Second, when the summands are indicators, biasing in the direction of the chosen summand is the same as conditioning on the event indicated by the chosen summand. Finally, when the summands, such as those in (57) or (58), are derived from an underlying process describing where every ball lands, such as $\mathbf{X} = (X_1, X_2, \ldots, X_b)$, then biasing the process of summands can be done by conditioning the entire underlying process on the event indicated by the chosen summand. We will find, for each summand, a coupling of $\mathbf{X}$ with $\mathbf{X}' = (\mathbf{X}$, conditional on the event indicated by that summand), which will give a coupling of the original $C(b,n)$ with the conditioned version $C'(b,n)$ such that $|C'(b,n) - C(b,n)|$ is bounded.

Consider the sum in (58). Assume $\mathbb{E}\, Y_{jk} > 0$. The event indicated by Y_{jk} is an intersection of two independent events, so conditioning on this is the same as conditioning on ball j landing in box k, and at least one of balls $1, 2, \ldots, j-1$ landing in box k. The sum $S = Z_{1k} + Z_{2k} + \cdots + Z_{j-1,k}$ is a sum of independent Bernoulli random variables, so by the sandwich principle [ArB, Cor. 7.1] there is a coupling of S with S' in which $S' - S \in \{0, 1\}$ for all outcomes ω, where S' is distributed as S conditioned to be nonzero.

This coupling of S with S' lifts to a coupling of $\mathbf{X}$ with $\mathbf{X}'$, in which $X_i = X'_i$ for all $i > j$, $X'_j = k$, and, either $S' - S = 0$ and $X_i = X'_i$ for $i = 1$ to $j-1$, or else $S' - S = 1$ and $X_i = X'_i$ for $i = 1$ to $j-1$ with a single exception I, with $X_I \neq k$, and $X'_I = k$. (To see this, begin with the observation that we have given values $p_1, \ldots, p_{j-1}$ with $p_i = \mathbb{E}\, Z_{ik} = \mathbb{P}(X_i = k)$, and there is a unique distribution for a permutation $(I_1, \ldots, I_{j-1})$ of $\{1, 2, \ldots, j-1\}$, such that starting from the all zero vector in $\{0,1\}^{j-1}$, and changing coordinates one at a time to one, according to the indices $I_1, \ldots, I_{j-1}$, yields the process $(Z_{ik}, Z_{2k}, \ldots, Z_{j-1,k})$ conditional on successively $S = 0, S = 1, \ldots, S = j-1$. Indeed this is explicitly the distribution of the size biased permutation of $(p_1, \ldots, p_{j-1})$.) For a summary of the changes in going from $\mathbf{X}$ and $C(b,n)$ to $\mathbf{X}'$ and $C'(b,n)$, ball j might move to box k, causing

C to change by -1, 0, or 1 (i.e., maybe lose a collision in the box X_j where ball j used to land, maybe gain a collision in box k) and there might also be one ball, with random label I in the range 1 to $j-1$, which moves to box k, causing an additional change to C by 0 or 1. (Minus 1 is not a possibility, since ball I, upon moving to box k, causes at least one additional collision.) The net result is that our coupling has $C'(b,n) - C(b,n) \in \{-1,0,1,2\}$; we have a 2-bounded size bias coupling. A general principle relating bounded couplings, monotone couplings, and bounded monotone couplings, [ArB, Prop. 7.1], now implies that there exists a coupling of $C(b,n)$ with its size biased version $C'(b,n)$, for which $0 \le C' - C \le 2$. The conclusion $C'(b,n) - C(b,n) \in \{0,1,2\}$ follows since both C' and C are integer valued.

Now consider the sum in (57), *and* assume that we are in the classical occupancy problem, i.e., that each X_i is *uniformly* distributed on the boxes $1,\dots,n$. The event indicated by the summand W_j may be expressed as

$$W_j = 1(S>0) \text{ where } S := \sum_{i=1}^{j-1} 1(X_i = X_j).$$

Thanks to the uniform distributions of the $X_1,\dots,X_j$, the distribution of S is Binomial$(j-1, \frac{1}{n})$. As in the previous paragraph, we couple S to S', distributed as S conditional on being strictly positive, by adding either 0 or 1, and this lifts to a coupling of $\mathbf{X}$ with $\mathbf{X}'$ in which either no ball moves, or else exactly one ball, with random index I, moves from a box other than X_j, to box X_j, where it causes one additional collision. We have $C' - C \in \{0,1\}$ for all outcomes, i.e., we have a 1-bounded monotone coupling. □

In the setting considered in this paper, the X_i's are uniformly distributed on $1,\dots,n$, so the proposition provides a 1-bounded monotone coupling of C' with C. Combining this with the main result of [GhG] immediately gives, with $\mu := \mathbb{E}\, C(b,n)$,

$$\begin{aligned} \mathbb{P}(C(b,n) - \mu \le -t) &\le \exp(-t^2/(2\mu)), \\ \mathbb{P}(C(b,n) - \mu \ge t) &\le \exp(-t^2/(2\mu + t)). \end{aligned} \tag{59}$$

for all $t > 0$ and all b, n. This strengthens the Azuma bounds from (56).

7.3. Uniform integrability for $(B(c,n) - \beta(c,n))/\sqrt{n}$.

Lemma 16. *Assume, as in Theorem 14, that we are given positive integers $c_1, c_2, \dots$ with $\lim_{n\to\infty} c_n = \infty$ and $\lim_{n\to\infty} c_n/n = \alpha_0 \in [0,\infty)$. With $\beta(c_n, n)$ given by (46), there exists $n_0 < \infty$ and $\epsilon > 0$ such that for all $n \ge n_0$ and for all y,*

$$\mathbb{P}(|B(c,n) - \beta(c,n)| \ge y\sqrt{n}) \le \exp(-\min(y, y^2)/104\,).$$

Proof. We check the bound for $\mathbb{P}(B(c,n) - \beta(c,n) \ge y\sqrt{n})$ with $y \ge 0$; the case of the other sign is comparatively easy and we omit the details. Start from (51) and (52), and write out explicitly $\mathbb{E}\, N_0 \equiv \mathbb{E}\, N_0(b,n) = n\,(1 - \frac{1}{n})^b$. This yields

$$\mathbb{P}(B(c,n) > b) = \mathbb{P}(N_0(b,n) - \mathbb{E}\, N_0 < -t) = \mathbb{P}(C(b,n) - \mathbb{E}\, C(b,n) < -t)$$

where

$$t = n(1 - \frac{1}{n})^b - ne^{-\lambda} + y\sqrt{n}. \tag{60}$$

The precise goal is to show that there exist n_0, y_0 such that for all $n > n_0, y > y_0$, $t^2/\mathbb{E}\,C(b,n) \geq \ell(y)$ with $\ell(y)/\log y \to \infty$ as $y \to \infty$. That is, we want a lower bound on $t^2/\mathbb{E}\,C(b,n)$ that grows with y, and is uniform in c, n. The analysis is similar to that in the proof of Theorem 14, but we want an inequality, carefully processed to show uniformity.

Using $-\log(1-\frac{1}{n}) = \frac{1}{n} + \frac{1}{2n^2} + \frac{1}{3n^3} + \cdots \leq \frac{1}{n} + \frac{1}{n^2}$ for $n \geq 2$, and $e^z - 1 \geq z$ for all real z, we have, for $n \geq 2$,

$$\begin{aligned} n(1-\frac{1}{n})^b - ne^{-b/n} &\geq n\left(\exp\left(-b\left(1/n + 1/n^2\right)\right) - e^{-b/n}\right) \\ &= n\, e^{-b/n}(\exp(-b/n^2) - 1) \\ &\geq -n\, e^{-b/n}\, \frac{b}{n^2} \geq -\frac{b}{n} \end{aligned}$$

Using $e^z - 1 \geq z$ again, we have

$$n(e^{-b/n} - e^{-\lambda}) = ne^{-\lambda}\left(\exp(-y/\sqrt{n}) - 1\right) \geq -e^{-\lambda} y\sqrt{n}.$$

Adding these two bounds, together with the final term of t from (60), we have

$$t \geq -\frac{b}{n} + (1 - e^{-\lambda})y\sqrt{n} = -\lambda - y/\sqrt{n} + (1 - e^{-\lambda})y\sqrt{n}.$$

For the delicate case, which is $c \to \infty, c/n \to \alpha_0 = 0$, we have $\beta^2(c,n) \sim 2cn$ hence $\lambda \to 0$. (The case $\alpha_0 > 0$ so that $1 - e^{-\lambda} \to 1 - e^{-\lambda_0} > 0$, hence $t \asymp y\sqrt{n}$ is *very easy* in comparison, and can even be handled via Azuma-Hoeffding; we omit further details.) With some choice $n_0 \geq 16$, for all $n > n_0$, $1 - e^{-\lambda} > \lambda/2$, hence $t \geq -\lambda - y/\sqrt{n} + \frac{1}{2}\lambda y\sqrt{n}$, and hence for all $y \geq 1$ we have $t \geq -y/\sqrt{n} + \frac{1}{4}\lambda y\sqrt{n}$. Finally, since $n\lambda = \beta(c,n) \to \infty$, increasing n_0 if needed, for all $y \geq 1, n \geq n_0$, we have $t \geq \frac{1}{5}\lambda y\sqrt{n}$. Squared, $t^2 \geq \frac{1}{25}\lambda^2 n y^2$, and since $\lambda^2 n = \beta^2(c,n)/n \sim 2cn/n = 2c$, increasing n_0 if needed, for all $y \geq 1, n \geq n_0$, we have $t^2 \geq \frac{c}{13}y^2$.

The upper bound (59) has the form: for $t \geq 0$, $\mathbb{P}(C(b,n) - \mathbb{E}\,C(b,n) \leq -t) \leq \exp(-r)$ where $r := t^2/(2\,\mathbb{E}\,C(b,n))$. We are in good shape when $\mathbb{E}\,C(b,n) \leq 4c$, which yields $r \geq y^2/104$. (Essentially, this is the main range, with $y^2 = O(c)$.)

For the remaining cases, where y is so large that $\mathbb{E}\,C(b,n) > 4c$, we bypass (52) and work directly with (51) and the duality. Write $\mu := \mathbb{E}\,C(b,n)$. With $y \geq 0$ and $b = \beta(c,n) + y\sqrt{n}$ such that $\mu > 4c$, hence $t := \mu - c > \frac{1}{2}\mu$,

$$\begin{aligned} \mathbb{P}(B(c,n) > b) &= \mathbb{P}(C(b,n) < c) = \mathbb{P}(C(b,n) - \mu < c - \mu) \\ &\leq \mathbb{P}(C(b,n) - \mu < -\frac{1}{2}\mu) \leq \exp(-r) \text{ where } r = \frac{(\mu/2)^2}{2\mu} = \mu/8. \end{aligned}$$

Now in case $y \leq n^{2/5}$, using $\beta \sim \sqrt{2cn} = o(n)$, so uniformly in $y \leq n^{2/5}$, $b = o(n)$ and $4c < \mu \sim b^2/(2n) \sim (\sqrt{2c} + y)^2/2$, hence $\mu > 4c$ and $c \to \infty$ implies $\inf_y y^2/\mu \geq 1/2$ so for sufficiently large $n, \mathbb{P}(B(c,n) > b) \leq \exp(-y^2/17)$. Also, in case $y \geq n^{3/5}$, we have $b > n^{6/5}$ and $B(c,n) \leq c + n$, hence, for sufficiently large n, $\mathbb{P}(B(c,n) > b) = 0$. To cover the missing range, if $n^{2/5} \leq y \leq n^{3/5}$ we simply use $y' = \sqrt{y}$ and $\mathbb{P}(B(c,n) \geq \beta + y\sqrt{n}) \leq \mathbb{P}(B(c,n) \geq \beta + y'\sqrt{n})$. □

Lemma 17. *Assume, as in Theorem 14, that we are given positive integers $c_1, c_2, \ldots$ with $\lim_{n\to\infty} c_n = \infty$ and $\lim_{n\to\infty} c_n/n = \alpha_0 \in [0, \infty)$. With $\beta(c_n, n)$ given by (46),*

there exists $n_0 < \infty$, such that for every $k = 1, 2, \ldots$, the family

$$\left\{ \left(\frac{B(c_n, n) - \beta(c_n, n)}{\sqrt{n}} \right)^k : n \geq n_0 \right\}$$

is uniformly integrable.

Proof. As in the proof of Lemma 10, the uniform and super-polynomial decaying upper bound from Lemma 16 implies uniform boundedness of the $(k+1)$-st moments for the $(B(c_n, n) - \beta(c_n, n))/\sqrt{n})$, $n \geq n_0$, which in turn implies uniform integrability of the family of k-th powers. □

8. Moments and variance in the centered case

Corollary 18. *As in Theorem 14, suppose* $\lim_{n\to\infty} c_n = \infty$ *and* $\lim_{n\to\infty} c_n/n = \alpha_0 \in [0, \infty)$. *Then, with w and g as defined by (45) and (48), and with $\lambda_0 = w^{-1}(\alpha_0)$,*

$$\operatorname{Var} B(c_n, n) \sim n\, g(\lambda_0)^2.$$

In particular, if $c_n \to \infty$ with $c_n = o(n)$, then $\operatorname{Var} B(c_n, n) \sim n/2$. *Furthermore,*

$$\begin{array}{rcll} \mathbb{E}[(B(c_n, n) - \beta(c_n, n))^k] & = & o(n^{k/2}) & \text{for } k = 1, 3, 5, \ldots \\ \mathbb{E}[(B(c_n, n) - \beta(c_n, n))^k] & \sim & (k-1)!!\, g(\lambda_0)^k\, n^{k/2} & \text{for } k = 2, 4, 6, \ldots \end{array}$$

In the display, $(k-1)!! = (k-1)(k-3)\cdots(5)(3)(1)$.

Proof. These claims follow from the distributional limit in Theorem 14, together with the uniform integrability from Lemma 17, and the moments of the standard normal. □

For the reader's convenience, we re-formulate some of our results for $c = o(n)$:

Corollary 19. *If $c = o(n)$, then*

$$\mathbb{E}B(c_n, n) \sim \gamma(c)\sqrt{2cn} \quad \text{and} \quad \operatorname{Var} B(c_n, n) \sim 2c(1 - \gamma(c)^2)\, n$$

for γ as in (36).

Proof. Combine Corollary 18 with equations (37) and (40) from Corollary 13. □

The claim for the expectation in Corollaries 18 and 19 extends [KuS, Th. 1] from the regime $c = o(n^{1/4})$ to $c = O(n)$. Furthermore, when $c = o(n)$, we have:

$$\lim_{n\to\infty} \frac{\sqrt{\operatorname{Var} B(c, n)}}{\mathbb{E}B(c, n)} = \frac{\sqrt{1 - \gamma(c)^2}}{\gamma(c)} = \frac{1}{2\sqrt{c}} + \frac{1}{32\sqrt{c}^3} - \frac{9}{1024\sqrt{c}^5} + \cdots$$

This justifies the following claim made in [KuS, p. 221]: "It turns out that the variance, when compared to the expected [value], is relatively low, especially if the number [c] ... is not too small."

Acknowledgements. We thank the referee for their insightful comments. SG's research was partially supported by NSF grant DMS-1201542, Emory University, and the Charles T. Winship Fund.

References

[AAR] G.E. Andrews, R. Askey, and R. Roy, *Special functions*, Encyclopedia of Mathematics and its Applications, vol. 71, Cambridge University Press, 1999.

[ArB] R. Arratia and P. Baxendale, *Bounded size bias coupling: a gamma function bound, and universal Dickman-function behavior*, to appear in *Probability Theory and Related Fields*, doi:10.1007/s00440-014-0572-x.

[ArGG] Richard Arratia, Larry Goldstein, and Louis Gordon, *Poisson approximation and the Chen-Stein method*, Statist. Sci. **5** (1990), no. 4, 403–434.

[ArGK] R. Arratia, L. Goldstein, and F. Kochman, *Size bias for one and all*, 2013, arXiv:1308.2729.

[BG] Jay Bartroff and Larry Goldstein, *A Berry-Esseen bound for the uniform multinomial occupancy model*, Electron. J. Probab. **18** (2013), no. 27, 29.

[BGI] J. Bartroff, L. Goldstein, and Ü. Işlak, *Bounded size biased couplings for log concave distributions and concentration of measure for occupancy models*, 2013, preprint.

[BHJ] A. D. Barbour, Lars Holst, and Svante Janson, *Poisson approximation*, Oxford Studies in Probability, vol. 2, The Clarendon Press Oxford University Press, New York, 1992, Oxford Science Publications.

[CaP] Michael Camarri and Jim Pitman, *Limit distributions and random trees derived from the birthday problem with unequal probabilities*, Electron. J. Probab. **5** (2000), no. 2, 18 pp.

[Cor$^+$] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and D.E. Knuth, *On the Lambert W function*, Adv. Comp. Math. **5** (1996), 329–359.

[D] Rick Durrett, *Probability: theory and examples*, 4th ed., Cambridge University Press, 2010.

[E] Gunnar Englund, *A remainder term estimate for the normal approximation of classical occupancy*, Annals of Probability **9** (1981), no. 4, 684–692.

[F] W. Feller, *An introduction to probability theory and its applications*, 3rd ed., vol. 1, Wiley, New York, 1968.

[GhG] Subhankar Ghosh and Larry Goldstein, *Concentration of measures via size-biased couplings*, Probab. Theory Related Fields **149** (2011), no. 1-2, 271–278.

[GnHP] Alexander Gnedin, Ben Hansen, and Jim Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws*, Probability Surveys **4** (2007), 146–171.

[H 86] Lars Holst, *On birthday, collectors', occupancy and other classical urn problems*, International Statistics Review (1986), 15–27.

[H 95] ———, *The general birthday problem*, Random Structures Algorithms **6** (1995), no. 2-3, 201–208.

[JK] Norman L. Johnson and Samuel Kotz, *Urn models and their application*, Wiley, 1977.

[JKB] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan, *Continuous univariate distributions*, 2nd ed., vol. 1, Wiley, 1994.

[Kn 2] D.E. Knuth, *The art of computer programming: seminumerical algorithms*, 2nd ed., vol. 2, Addison-Wesley, 1981.

[Kn 3] ———, *The art of computer programming: sorting and searching*, 2nd ed., vol. 3, Addison-Wesley, 1998.

[KoSC] V.F. Kolchin, B.A. Sevastyanov, and V.P. Chistyakov, *Random allocations*, Wiley, 1978.

[KuS] F. Kuhn and R. Struik, *Random walks revisited: extensions of Pollard's Rho algorithm for computing multiple discrete logarithms*, Selected areas in cryptology (SAC 2001, Toronto, ON) (S. Vaudenay and A. Youssef, eds.), Lecture Notes in Computer Science, vol. 2259, Springer, 2001, pp. 212–229.

[Mi] V. G. Mikhaĭlov, *The central limit theorem for a scheme of independent allocation of particles by cells*, Trudy Mat. Inst. Steklov. **157** (1981), 138–152, 236, Number theory, mathematical analysis and their applications.

[Mo] F. Mosteller, *Fifty challenging problems in probability with solutions*, Dover, 1987, reprint of the 1965 Addision-Wesley edition.

[R] A. Rényi, *Three new proofs and a generalization of a theorem of Irving Weiss*, Publ. Math. Inst. Hung. Acad. Sci. **7** (1962), 203–214, [Reprinted in vol. III of his *Selected Papers*, pp. 67–77].

[Slo] N.J.A. Sloane, *The on-line encyclopedia of integer sequences*, available at `oeis.org`.

[St] Steven Strogatz, *It's my birthday too, yeah*, from the New York Times Opinionator blog at `http://opinionator.blogs.nytimes.com/2012/10/01/its-my-birthday-too-yeah/`, October 2012.

[T] Hale F. Trotter, *Eigenvalue distributions of large Hermitian matrices; Wigner's semi-circle law and a theorem of Kac, Murdock, and Szegő*, Adv. in Math. **54** (1984), no. 1, 67–82.

[W] Irving Weiss, *Limiting distributions in some occupancy problems*, Ann. Math. Statist. **29** (1958), 874–884.

Arratia: Department of Mathematics, University of Southern California, Department of Mathematics, 3620 S. Vermont Ave., KAP 104, Los Angeles, CA 90089-2532

E-mail address: `rarratia at usc.edu`

Garibaldi: Institute for Pure and Applied Mathematics, UCLA, 460 Portola Plaza, Box 957121, Los Angeles, California 90095-7121, USA

E-mail address: `skip at member.ams.org`

Kilian: IDA Center for Communications Research, 805 Bunn Dr., Princeton, NJ 08540