Independent clocks for infinite coupon collection

Jacob Richey with GPT-6-Astra

October 25, 2026

Abstract

Independent clocks can describe which coupons have appeared by a late time and still fail completely to describe their future discoveries. For iid sampling with probabilities proportional to j^{-2}, the occupied set is asymptotically indistinguishable from the independent model, while simultaneous arrivals distinguish their future histories with probability one. With probabilities proportional to j^{-3}, even the entire future can be approximated. We explain the classical fixed-time argument and derive the criterion separating these behaviors: the decreasingly ordered probabilities must satisfy $\sum_j jp_j < \infty$. A stretched exponential example gives an error of order $(\log n)^2/n$ and a worked Gaussian limit for a weighted coupon count.

1 Introduction

Suppose that we draw coupons one at a time, independently, with coupon j having probability proportional to j^{-2}, for $j = 1, 2, \ldots$. Each coupon has a geometric waiting time until its first appearance. If we give the coupons independent clocks with these same waiting-time laws, how much of the collection process do we preserve? At a fixed late time, the entire lists of discovered coupons can be coupled to agree with probability tending to one. Yet if we keep watching, the independent clocks ring in pairs or larger groups infinitely often, almost surely. The collector can never discover two coupons on the same draw. The lists at a single time become indistinguishable in law, while their future histories remain distinguishable with certainty.

Changing the probabilities to be proportional to j^{-3} changes the answer for the future: given a sufficiently late starting time, the two processes can be coupled so that all subsequent discoveries agree, with probability close to one. More generally, for probabilities proportional to j^{-a}, $a > 1$, this is possible exactly when $a > 2$. The fixed-time comparison works for every $a > 1$, and indeed for every fixed countable sampling distribution. The difference is whether the errors from treating rare discoveries as independent can accumulate over the whole future.

This note develops that comparison by elementary coupling and collision estimates. The starting point is the classical observation that Poissonizing the number of draws makes the coupon counts independent. At a fixed large number of draws, the dependence can still be removed approximately by separating the coupons that have almost certainly appeared from the rare ones. Keeping the future requires controlling the cost of maintaining this independence at every subsequent step. We first state the two conclusions precisely, then explain their relationship to the occupancy literature.

Let $X_1, X_2, \ldots$ be iid with a fixed probability mass function $(p_j)_{j\in J}$ on a finite or countably infinite set J, with all $p_j > 0$, and write

$$V_n = \{X_1, \ldots, X_n\}, \qquad V_0 = \varnothing.$$

Take independent geometric random variables G_j with

$$\mathbb{P}(G_j > n) = (1 - p_j)^n, \qquad n = 0, 1, \ldots, \qquad W_n = \{j : G_j \le n\}.$$

Each label has the same probability of belonging to V_n and W_n, but the membership indicators in W_n are independent. These sets are finite almost surely, since $\mathbb{E}|W_n| \le \sum_j np_j = n$, and are nested as n increases. Thus one collection of clocks supplies the comparison at every time.

Write $d_{\mathrm{TV}}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$ for total variation distance, and put

$$\epsilon_n = d_{\mathrm{TV}}(\mathcal{L}(V_n), \mathcal{L}(W_n)),$$
$$D_n = d_{\mathrm{TV}}(\mathcal{L}((V_t)_{t\ge n}), \mathcal{L}((W_t)_{t\ge n})).$$

Total variation lies in $[0,1]$ and is the smallest possible disagreement probability in a coupling. The path laws use the product sigma field on sequences of finite subsets of J. In D_n, the starting occupied sets and every later discovery are compared at their original integer times.

Theorem 1.1. *For every fixed countable probability mass function, $\epsilon_n \to 0$. Arrange its positive probabilities in decreasing order, $p_1 \ge p_2 \ge \cdots$. Then*

$$D_n \longrightarrow 0 \quad \Longleftrightarrow \quad \sum_{j\ge 1} jp_j < \infty.$$

If the sum diverges, $D_n = 1$ for every $n \ge 1$. For finite support the sum is finite.

The first assertion is a consequence of classical Poisson approximation. The useful estimate is particularly explicit in Barbour and Gnedin (2009) [2, Lemma 2.1, equation (2.6)]: for a tail of labels of total mass q, their counts in n draws are within q, in total variation, of independent Poisson counts with means np_j. Their proof combines this estimate with truncation of the frequent labels. Barbour (2009) [1, Section 1] explicitly carries the comparison over to the *labeled occupancy indicators*. To obtain the exact one-coupon marginals used here, compare independent binomial counts to the same Poisson variables; this costs at most another q. Proposition 2.1 records the resulting bound. Barbour and Gnedin credit Le Cam's Poisson approximation (1960) [9] and Michel's compound Poisson bound (1987) [10]; our proof uses the scalar bound of Barbour and Hall (1984) [3].

There are also limit theorems for the late discoveries themselves. For example, Derbazi, Gnedin and Marynych (2024) [4, Theorems 7.1 and 8.4] imply the following for $p_j \propto j^{-a}$, $a > 1$. If

$$s_n = \frac{an}{\sum_j (1 - e^{-np_j})},$$

then

$$\left(|V_{n+\lfloor s_n t\rfloor}| - |V_n|\right)_{t\ge 0} \xrightarrow{d} (\Pi_t)_{t\ge 0}, \tag{1.1}$$

where Π is a unit-rate Poisson process and convergence is in the Skorokhod J_1 topology. This holds on both sides of the threshold $a = 2$. It describes discovery times after a local rescaling and with the labels forgotten. In contrast, D_n compares the labeled sets at every original time, over an unbounded future. Thus a Poisson limit such as (1.1) is compatible with $D_n = 1$. Their de-Poissonization argument also uses the Barbour–Gnedin estimate above.

For this stronger comparison of labeled histories, a related coupling appears in Eriksson's 2025 preprint, revised in 2026 [5, proof of Theorem 14]. In his model, edges sampled from a fixed distribution introduce their endpoints into a growing graph. In continuous time, if the graph is eventually connected forever, he couples its vertex set with the occupied set generated by independent exponential clocks. The sets agree from some finite random time onward. Simultaneous births create discrepancies; controlling their number allows the processes to agree eventually. This gives a precedent for the coupling mechanism in a different model. Here a sum of one-step errors gives the sufficient condition in Theorem 1.1, and infinitely many simultaneous arrivals in the independent

model give the converse. We do not know a reference for this exact discrete-time criterion. We present its elementary proof without a claim of priority.

Once the occupied set has been replaced by independent indicators, its statistics become easier to study. This is a longstanding use of Poissonization in infinite occupancy theory, going back to Karlin's central limit theorems (1967) [8]; see Gnedin, Hansen and Pitman (2007) [6] for a survey. Section 4 works through $p_v \propto e^{-\alpha\sqrt{v}}$ and the weighted count $\sum_{v\in V_n}\sqrt{v}$. Labels with probability comparable to $1/n$ supply both the random fluctuations of this count and the clock collisions that determine the future approximation error. The example therefore lets us see the same comparison at work in a limit theorem and in an error bound.

1.1 Overview of the proof

The first step is to locate the randomness that remains at a late time. Divide the coupons into a finite set H of frequent labels and its complement. With high probability all labels in H have appeared in both models. It therefore suffices to compare the remaining counts, whose total probability mass is $q = \sum_{j\notin H} p_j$. Their total number of occurrences in the collector is $\mathrm{Bin}(n, q)$; conditional on this number, each occurrence receives label j with probability p_j/q. Replacing the total by $\mathrm{Poi}(nq)$ makes the labeled counts independent Poisson variables and costs at most q in total variation. Independent binomial counts with the correct marginals are also within q of those Poisson variables. Passing to nonzero indicators gives

$$\epsilon_n \le 2\sum_{j\in H}(1-p_j)^n + 2\sum_{j\notin H} p_j.$$

First let n grow with H fixed, then enlarge H. This is why no tail assumption is needed for the first assertion of the theorem.

To keep the processes together after time n, start with a coupling of V_n and W_n, then couple their next steps whenever they still agree. If U is the current set of unseen labels, the collector adds label j with probability p_j and adds at most one label. By memorylessness, the clock process adds each $j \in U$ independently with probability p_j. The discrepancy between these transition laws is at most $2\sum_{i<j,\ i,j\in U} p_i p_j$. A pair i, j is still unseen in the collector at time t with probability $(1-p_i-p_j)^t$. Summing over all possible failure times therefore gives

$$D_n \le \epsilon_n + 2\sum_{i<j}\frac{p_i p_j}{p_i+p_j}(1-p_i-p_j)^n.$$

In decreasing probability order, $p_i p_j/(p_i+p_j)$ lies between $p_j/2$ and p_j. The sum is consequently dominated by a summable series when $\sum_j jp_j < \infty$, and then tends to zero.

The converse uses the collisions that the collector cannot have. Two independent clocks satisfy

$$\mathbb{P}(G_i = G_j) = \frac{p_i p_j}{p_i + p_j - p_i p_j},$$

again between $p_j/2$ and p_j for $i < j$. Divergence of $\sum_j jp_j$ thus gives infinitely many collisions in expectation, but an almost sure statement is needed to distinguish the path laws. If C_N counts colliding pairs up to time N, disjoint pairs give independent indicators, and overlapping pairs can be controlled by triple collisions. This yields $\mathrm{Var}\, C_N \le 3\mathbb{E}C_N$. Chebyshev's inequality then forces $C_N \to \infty$ almost surely when its mean diverges. There are only finitely many arrivals at each fixed time, so multiple arrivals occur arbitrarily late. Their presence separates the two future path laws.

2 Approximation at a fixed time

For a finite set $H \subseteq J$, define

$$h_n(H) = \sum_{j \in H} (1 - p_j)^n, \qquad q(H) = \sum_{j \notin H} p_j.$$

The first quantity bounds the probability that some coupon of H is missing; the second is the mass of the remaining labels. Keeping both terms in the estimate will also give a rate in the example below.

Proposition 2.1 (Classical truncation bound)**.** *For every $n \geq 1$ and finite $H \subseteq J$,*

$$\epsilon_n \leq 2h_n(H) + 2q(H). \tag{2.1}$$

Proof. The binomial case of the Poisson approximation bound of Barbour and Hall [3] gives, for $q \in [0, 1]$,

$$d_{\mathrm{TV}}(\mathrm{Bin}(n, q), \mathrm{Poi}(nq)) \leq (1 - e^{-nq})q \leq q. \tag{2.2}$$

The sets V_n and $V_n \cup H$ differ with probability at most $h_n(H)$, and the same holds for W_n and $W_n \cup H$. It remains to compare their restrictions to $J \setminus H$.

Write $q = q(H)$; the case $q = 0$ is immediate. Let $N_j = \sum_{i=1}^{n} \mathbf{1}_{\{X_i = j\}}$. The total $\sum_{j \notin H} N_j$ has law $\mathrm{Bin}(n, q)$. Conditional on this total being m, its m labels are iid with probabilities p_j/q. Independent variables $Z_j \sim \mathrm{Poi}(np_j)$, $j \notin H$, have a $\mathrm{Poi}(nq)$ total and the same conditional allocation. Coupling the totals using (2.2) and using the same labels when they agree gives

$$d_{\mathrm{TV}}(\mathcal{L}((N_j)_{j \notin H}), \mathcal{L}((Z_j)_{j \notin H})) \leq q.$$

Both totals are finite almost surely, so this construction applies to countably many labels.

Let $B_j \sim \mathrm{Bin}(n, p_j)$, $j \notin H$, be independent. Taking independent maximal couplings of (B_j, Z_j), (2.2) and the countable union bound give

$$d_{\mathrm{TV}}(\mathcal{L}((B_j)_{j \notin H}), \mathcal{L}((Z_j)_{j \notin H})) \leq \sum_{j \notin H} p_j = q.$$

The nonzero indicators of (B_j) have the law of the corresponding indicators in W_n. Mapping all counts to their nonzero indicators and using the triangle inequality proves (2.1). □

For fixed H, $h_n(H) \to 0$, so $\limsup_n \epsilon_n \leq 2q(H)$. Exhausting J by finite sets proves the first assertion of Theorem 1.1. Moreover, for every measurable statistic F_n of a finite set,

$$d_{\mathrm{TV}}(\mathcal{L}(F_n(V_n)), \mathcal{L}(F_n(W_n))) \leq \epsilon_n. \tag{2.3}$$

This includes statistics involving many labels or assigning unequal weights to them. Weak limits can therefore be calculated in the independent model and transferred to the collector. For unbounded statistics, moment asymptotics need a separate argument; we give one for the weighted count in Section 4.

3 The future discovery process

A comparison of the occupied sets supplies the initial coupling. To extend it through future discoveries, we need to compare the two transition laws at a state where the processes agree. Memorylessness of the geometric clocks makes this a calculation involving only the unseen labels.

Lemma 3.1. *For $U \subseteq J$, let K_U be the law that assigns probability p_i to $\{i\}$ for $i \in U$ and probability $1 - \sum_{i\in U} p_i$ to $\varnothing$. Let Q_U be the law of the random subset of U that includes each i independently with probability p_i. Then*

$$\delta(U) := d_{\mathrm{TV}}(K_U, Q_U) = \sum_{i\in U} p_i \left[1 - \prod_{j\in U\setminus\{i\}} (1-p_j)\right] \le 2 \sum_{\substack{i<j \\ i,j\in U}} p_i p_j. \tag{3.1}$$

Proof. A set sampled from Q_U is finite almost surely because its expected cardinality is $\sum_{i\in U} p_i \le 1$. Its singleton probabilities are $p_i \prod_{j\in U\setminus\{i\}}(1-p_j) \le p_i$, while the union bound gives $Q_U(\varnothing) \ge 1 - \sum_{i\in U} p_i = K_U(\varnothing)$. The positive part of $K_U - Q_U$ is therefore supported on singletons, and its total mass is the displayed sum. Applying the union bound to each bracket proves the inequality. □

Proposition 3.2. *For every $n \ge 1$,*

$$D_n \le \epsilon_n + 2\sum_{i<j} \frac{p_i p_j}{p_i + p_j}(1 - p_i - p_j)^n. \tag{3.2}$$

The sum may be infinite.

Proof. Take a maximal coupling at time n. Whenever the processes have agreed through time $t \ge n$, use a maximal coupling of their next-step laws; after their first disagreement, use independent transitions. This defines a coupling of the two Markov chains on the countable space of finite subsets of J. Let E_t be the event that they agree at all times from n through t. By Lemma 3.1,

$$\mathbb{P}(E_t \setminus E_{t+1}) = \mathbb{E}[\mathbf{1}_{E_t}\delta(J \setminus V_t)] \le 2\sum_{i<j} p_i p_j \mathbb{P}(i,j \notin V_t) = 2\sum_{i<j} p_i p_j (1 - p_i - p_j)^t.$$

Summing the disjoint first-disagreement events and applying Tonelli's theorem gives

$$\mathbb{P}(\text{some disagreement at a time } t \ge n) \le \epsilon_n + 2\sum_{i<j} p_i p_j \sum_{t=n}^{\infty} (1 - p_i - p_j)^t.$$

The geometric series and the coupling bound for total variation yield (3.2). □

The same pairs that appear in this upper bound determine whether the independent process has infinitely many simultaneous arrivals. The next lemma supplies the almost sure statement needed for the converse.

Lemma 3.3. *Let*

$$C_N = \sum_{i<j} \mathbf{1}_{\{G_i = G_j \le N\}}, \qquad C_\infty = \lim_{N\to\infty} C_N.$$

Then

$$\mathbb{E}C_\infty = \sum_{i<j} \frac{p_i p_j}{p_i + p_j - p_i p_j}, \qquad \operatorname{Var} C_N \le 3\mathbb{E}C_N < \infty. \tag{3.3}$$

If $\mathbb{E}C_\infty = \infty$, then $C_\infty = \infty$ almost surely, and collisions occur at arbitrarily late times.

Proof. Independence of G_i, G_j and summation of a geometric series give

$$\mathbb{P}(G_i = G_j) = \sum_{t\geq 1} p_i p_j [(1-p_i)(1-p_j)]^{t-1} = \frac{p_i p_j}{p_i + p_j - p_i p_j}.$$

Tonelli's theorem proves the expectation formula.

Put $a_{i,t} = p_i(1-p_i)^{t-1}$ and $b_t = \sum_i a_{i,t} \leq 1$. For a finite set of labels, indicators for pairs with disjoint label sets are independent. The covariance of overlapping pair indicators is at most the probability of a triple collision. If $C_N^{(m)}$ restricts the pair count to the first m labels, then

$$\begin{aligned}\operatorname{Var} C_N^{(m)} &\leq \mathbb{E}C_N^{(m)} + 6\sum_{t=1}^{N} \sum_{i<j<k\leq m} a_{i,t}a_{j,t}a_{k,t} \\ &\leq \mathbb{E}C_N^{(m)} + 2\sum_{t=1}^{N} b_t \sum_{i<j\leq m} a_{i,t}a_{j,t} \leq 3\mathbb{E}C_N^{(m)}.\end{aligned}$$

Here the middle inequality follows by expanding $(\sum_{i\leq m} a_{i,t}) \sum_{i<j\leq m} a_{i,t}a_{j,t}$, in which every triple product occurs three times. Also $\mathbb{E}C_N^{(m)} \leq \frac{1}{2}\sum_{t=1}^{N} b_t^2 \leq N/2$. Consequently $\mathbb{E}[(C_N^{(m)})^2] \leq 3N/2 + N^2/4$. Monotone convergence for the counts and their squares permits $m \to \infty$, proving the variance bound in (3.3).

If $\mu_N := \mathbb{E}C_N \to \infty$, then for each fixed M and $\mu_N > M$,

$$\mathbb{P}(C_N \leq M) \leq \frac{3\mu_N}{(\mu_N - M)^2} \longrightarrow 0.$$

Since C_N is increasing, $\mathbb{P}(C_\infty \leq M) = 0$ for every finite M. At each fixed time t, the number of arrivals has expectation $b_t \leq 1$ and is finite almost surely. Infinitely many colliding pairs therefore imply collisions at arbitrarily late times. □

Proof of Theorem 1.1. The assertion $\epsilon_n \to 0$ was proved after Proposition 2.1. In decreasing probability order, for $i < j$,

$$\frac{p_j}{2} \leq \frac{p_i p_j}{p_i + p_j} \leq p_j, \qquad \frac{p_j}{2} \leq \frac{p_i p_j}{p_i + p_j - p_i p_j} \leq p_j.$$

If $\sum_j jp_j < \infty$, the series in (3.2) is dominated by $\sum_j (j-1)p_j < \infty$. Each summand tends to zero, so dominated convergence and Proposition 3.2 give $D_n \to 0$.

If $\sum_j jp_j = \infty$, (3.3) diverges, and Lemma 3.3 gives simultaneous arrivals after every fixed time almost surely. For each n, the event that every transition after n adds at most one label has probability one for $(V_t)_{t\geq n}$ and zero for $(W_t)_{t\geq n}$. Thus $D_n = 1$. □

4 A stretched exponential collector

Fix $\alpha > 0$ and take

$$J = \{0, 1, \ldots\}, \qquad p_v = C_\alpha e^{-\alpha\sqrt{v}}, \qquad C_\alpha = \left(\sum_{v\geq 0} e^{-\alpha\sqrt{v}}\right)^{-1}. \tag{4.1}$$

These sampling probabilities have a stretched exponential tail; it is the discovery clocks that are geometric. Theorem 1.1 applies because $\sum_v (v+1)p_v < \infty$. We now calculate the approximation error and use independence to study

$$L_n = \sum_{v \in V_n} \sqrt{v}, \qquad \widetilde{L}_n = \sum_{v \in W_n} \sqrt{v}.$$

For this class of sampling probabilities, Iksanov, Kabluchko and Kotelnikova obtained functional Gaussian limits for occupancy counts in a 2021 preprint, published in 2022 [7, Theorem 1.1, Corollary 1.2 and Remark 1.3]. Their first generation is the ordinary infinite occupancy model. The weighted statistic here gives a direct example of the calculation after passing to independent indicators.

Write $t_n = \log(nC_\alpha)$, so that $np_v = e^{t_n - \alpha\sqrt{v}}$. The boundary between labels that have usually appeared and those that have usually not appeared lies near $\alpha\sqrt{v} = t_n$. An interval of fixed width in the coordinate $\alpha\sqrt{v}$, near this boundary, contains order t_n labels. Each has weight of order t_n, suggesting a variance of order t_n^3 for the weighted count. These labels also have clock collision probabilities of order $1/n$ per pair, suggesting an error of order t_n^2/n for the future history. The estimates below make both calculations precise.

4.1 Sums near the discovery boundary

The following estimates will be used for the coupling error and for the moments of $\widetilde{L}_n$. They express how a sum over labels becomes an integral over a bounded-width region around $\alpha\sqrt{v} = t_n$.

Lemma 4.1. *Set $q_{v,n} = 1 - (1-p_v)^n$. For each fixed nonnegative integer r, as $n \to \infty$,*

$$\sum_{v \geq 0} v^{r/2} q_{v,n}(1 - q_{v,n}) = \frac{2\log 2}{\alpha^{r+2}} t_n^{r+1} + O_{\alpha,r}(t_n^r), \tag{4.2}$$

and, for $k = 1, 2$,

$$\sum_{v \geq 0} v^{r/2} (np_v)^k e^{-np_v} = O_{\alpha,r}(t_n^{r+1}). \tag{4.3}$$

When $r = 0$, the weight at $v = 0$ is interpreted as one.

Proof. For an absolutely continuous function f with f, f' integrable on $[0, \infty)$, comparison on each interval $[v, v+1]$ gives

$$\left| \sum_{v \geq 0} f(v) - \int_0^\infty f(x)\, dx \right| \leq \int_0^\infty |f'(x)|\, dx. \tag{4.4}$$

Indeed, $|f(v) - f(x)| \leq \int_v^{v+1} |f'(s)|\, ds$ on that interval, and these bounds are summable.

Let $g(u) = e^{-u} - e^{-2u}$, and let F be any of $g(u), ue^{-u}, u^2e^{-u}$. Put $h_F(z) = F(e^{-z})$. For $z \geq 0$, $|h_F(z)| + |h_F'(z)| \leq Ke^{-z}$; for $z < 0$, this sum is at most $K(1 + e^{-3z})e^{-e^{-z}}$. Thus h_F and h_F' have finite absolute moments of every order. With $t = t_n$ and $y = \alpha\sqrt{x}$,

$$\int_0^\infty x^{r/2} F(nC_\alpha e^{-\alpha\sqrt{x}})\, dx = \frac{2}{\alpha^{r+2}} \int_0^\infty y^{r+1} h_F(y - t)\, dy. \tag{4.5}$$

After substituting $z = y - t$, expansion of $(t+z)^{r+1}$ and the moment bounds give

$$\int_0^\infty y^{r+1} h_F(y - t)\, dy = t^{r+1} \int_{\mathbb{R}} h_F(z)\, dz + O_r(t^r).$$

For example, the omitted leading-term tail is bounded by $t^{r+1}\int_{z<-t}|h_F(z)|\,dz = O(t^r)$, using its first absolute moment; the remaining terms are bounded by the higher moments. The variation of the function on the left of (4.5) is at most

$$\alpha^{-r}\int_0^\infty (ry^{r-1}|h_F(y-t)| + y^r|h_F'(y-t)|)\,dy = O_{\alpha,r}(t^r),$$

with the first term omitted for $r=0$. Equations (4.4)–(4.5) prove (4.3) and the corresponding asymptotic formula for $\sum_v v^{r/2}g(np_v)$, since

$$\int_{\mathbb{R}} g(e^{-z})\,dz = \int_0^\infty \frac{e^{-u}-e^{-2u}}{u}\,du = \log 2.$$

The last equality follows from $(e^{-u}-e^{-2u})/u = \int_1^2 e^{-su}\,ds$ and Tonelli's theorem.

Since $p_v \le C_\alpha < 1$, expansion of $-\log(1-p)$ gives

$$0 \le e^{-np} - (1-p)^n \le \frac{np^2}{2(1-C_\alpha)}e^{-np}, \qquad 0 \le p \le C_\alpha. \tag{4.6}$$

The map $a \mapsto a(1-a)$ is 1-Lipschitz on $[0,1]$. Consequently the difference between the sum in (4.2) and $\sum_v v^{r/2}g(np_v)$ is at most

$$\frac{1}{2(1-C_\alpha)n}\sum_v v^{r/2}(np_v)^2e^{-np_v} = O_{\alpha,r}(t_n^{r+1}/n) = o(t_n^r).$$

This proves (4.2). □

4.2 The approximation error

The truncation bound requires a cutoff beyond which the remaining probability mass is small. For the future history, we must also sum the transition errors. Both bounds have the same order in this example, and late clock collisions give a matching lower bound for the future history.

Proposition 4.2. *For the probabilities* (4.1),

$$\epsilon_n = O_\alpha\left(\frac{(\log n)^2}{n}\right), \qquad D_n = \Theta_\alpha\left(\frac{(\log n)^2}{n}\right).$$

Proof. For n sufficiently large, let $H_n = \{v : np_v \ge 3\log n\}$ and $b = \log(nC_\alpha/(3\log n)) > 0$. Then $|H_n| \le 1 + (b/\alpha)^2$ and $h_n(H_n) \le |H_n|n^{-3}$. Decreasing integral comparison gives

$$\sum_{\alpha\sqrt{v}>b} e^{-\alpha\sqrt{v}} \le \left(1 + \frac{2(b+1)}{\alpha^2}\right)e^{-b}.$$

It follows that $q(H_n) = O_\alpha((\log n)^2/n)$, and Proposition 2.1 proves the bound on ϵ_n.

The extra term in (3.2) is at most

$$2\sum_{t=n}^\infty \sum_{i<j} p_ip_je^{-t(p_i+p_j)} \le \sum_{t=n}^\infty \left(\sum_v p_ve^{-tp_v}\right)^2.$$

Lemma 4.1, with $r=0, k=1$, gives $\sum_v p_ve^{-tp_v} = O_\alpha(\log t/t)$. Integration of $(\log t)^2/t^2$ bounds the last display by $O_\alpha((\log n)^2/n)$, proving the upper bound on D_n.

For the lower bound, let

$$I_n = \{v : 1/n \le p_v \le 2/n\}, \qquad M_n = |I_n|,$$

and let Z_n count unordered pairs of labels in I_n whose clocks coincide at a time $n < t \le 2n$. The interval

$$\log(nC_\alpha/2) \le \alpha\sqrt{v} \le \log(nC_\alpha)$$

has $M_n = \Theta_\alpha(\log n)$ integer labels. Uniformly over distinct $i, j \in I_n$, for $n \ge 4$,

$$\frac{e^{-16}}{n} \le \sum_{t=n+1}^{2n} p_i p_j [(1-p_i)(1-p_j)]^{t-1} \le \frac{4}{n}.$$

The lower bound uses $\log(1-p) \ge -2p$ for $p \le 1/2$. Thus $\mathbb{E}Z_n = \Theta_\alpha(M_n^2/n)$. Disjoint pair events are independent; two overlapping pair events require a triple collision, whose probability in this interval is at most $n(2/n)^3 = 8/n^2$. Counting the disjoint and overlapping ordered pairs gives

$$\mathbb{E}[Z_n(Z_n-1)] = O\left(\frac{M_n^4 + M_n^3}{n^2}\right) = o(\mathbb{E}Z_n).$$

Since $\mathbf{1}_{\{z>0\}} \ge z - z(z-1)/2$ for nonnegative integers z, $\mathbb{P}(Z_n > 0) \ge \frac{1}{2}\mathbb{E}Z_n$ for all sufficiently large n. A multiple arrival in $(n, 2n]$ is measurable from the future history and has probability zero in the collector. Therefore $D_n \ge \mathbb{P}(Z_n > 0) = \Omega_\alpha((\log n)^2/n)$. □

4.3 A weighted Gaussian limit

The independent statistic is a sum of Bernoulli variables with unequal weights. Lemma 4.1 gives its variance and the third-moment bound needed for a central limit theorem. The mean needs one further term beyond its leading cubic order: its quadratic correction is larger than the fluctuation scale.

Proposition 4.3. *Let γ be Euler's constant and $\mu_n = \mathbb{E}L_n = \mathbb{E}\widetilde{L}_n$. As $n \to \infty$,*

$$\mu_n = \frac{2}{3\alpha^3}t_n^3 + \frac{2\gamma}{\alpha^3}t_n^2 + O_\alpha(t_n), \tag{4.7}$$

$$\operatorname{Var} L_n \sim \operatorname{Var} \widetilde{L}_n \sim \frac{2\log 2}{\alpha^4}t_n^3, \tag{4.8}$$

$$\frac{L_n - \mu_n}{(\sqrt{2\log 2}/\alpha^2)t_n^{3/2}} \xrightarrow{d} \mathcal{N}(0,1). \tag{4.9}$$

The centering in (4.9) may also be replaced by the two explicit terms in (4.7).

Proof. The one-coupon marginals give $\mu_n = \sum_v \sqrt{v}\,[1-(1-p_v)^n]$. By (4.6) and (4.3), replacing $(1-p_v)^n$ by e^{-np_v} changes this sum by $O_\alpha(t_n^2/n)$. Put $t = t_n$ and $f_t(x) = \sqrt{x}\,[1 - e^{-e^{t-\alpha\sqrt{x}}}]$. Its variation is bounded by

$$\alpha^{-1}\int_0^\infty \left[1 - e^{-e^{t-y}} + ye^{t-y}e^{-e^{t-y}}\right] dy = O_\alpha(t).$$

For the first term, splitting at $y = t$ and using $1 - e^{-u} \le u$ gives a bound $t+1$; the second is $O(t)$ by the moment estimates for h_F in Lemma 4.1. Equation (4.4) now yields

$$\mu_n = \frac{2}{\alpha^3}\int_0^\infty y^2(1 - e^{-e^{t-y}})\,dy + O_\alpha(t) = \frac{2}{3\alpha^3}\mathbb{E}[(t+Z)_+^3] + O_\alpha(t),$$

where Z has the standard Gumbel distribution $\mathbb{P}(Z \le z) = e^{-e^{-z}}$, and the last equality follows by Tonelli. Here $x_+ = \max\{x, 0\}$, $\mathbb{E}Z = \gamma$, and $\mathbb{E}|Z|^3 < \infty$. On $\{Z < -t\}$, $|t + Z| \le |Z|$, so

$$\mathbb{E}[(t+Z)_+^3] = \mathbb{E}[(t+Z)^3] + O(1) = t^3 + 3\gamma t^2 + O(t).$$

This proves (4.7).

Independence and (4.2) with $r = 2$ give $\operatorname{Var} \widetilde{L}_n = (2 \log 2/\alpha^4) t_n^3 + O_\alpha(t_n^2)$. For distinct labels, the collector indicators $A_v = \mathbf{1}_{\{v \in V_n\}}$ satisfy

$$\operatorname{Cov}(A_v, A_w) = (1 - p_v - p_w)^n - (1 - p_v)^n (1 - p_w)^n \le 0.$$

Using $a^n - b^n \le n(a-b)a^{n-1}$ for $0 \le b \le a \le 1$,

$$\begin{aligned} 0 \le \operatorname{Var} \widetilde{L}_n - \operatorname{Var} L_n &\le n \left(\sum_v \sqrt{v}\, p_v (1 - p_v)^{n-1} \right)^2 \\ &\le \frac{e^{2C_\alpha}}{n} \left(\sum_v \sqrt{v}\, (np_v) e^{-np_v} \right)^2 = O_\alpha(t_n^4/n) = o(1). \end{aligned}$$

The infinite sums follow by L^2 truncation: the collector satisfies $L_n \le \sum_{i=1}^n \sqrt{X_i}$, whose square is integrable, and the independent sum has finite mean and variance because $q_{v,n} \le np_v$ and $\sum_v v p_v < \infty$. This proves (4.8).

Let $B_{v,n} = \mathbf{1}_{\{v \in W_n\}}$. For a Bernoulli variable of mean q, the third absolute centered moment is at most $q(1-q)$. Hence (4.2) with $r = 3$ gives

$$\sum_v \mathbb{E}\left|\sqrt{v}\,(B_{v,n} - q_{v,n})\right|^3 \le \sum_v v^{3/2} q_{v,n}(1 - q_{v,n}) = O_\alpha(t_n^4).$$

Dividing by $(\operatorname{Var} \widetilde{L}_n)^{3/2}$ gives $O_\alpha(t_n^{-1/2})$, which tends to zero. In each row, truncate to finitely many labels so that the discarded variance is at most one. The Lyapunov central limit theorem applies to the truncated sums, and the centered discarded sums tend to zero in L^2 after division by $t_n^{3/2}$. Thus (4.9) holds for $\widetilde{L}_n$. Equation (2.3) transfers it to L_n. Finally, the $O_\alpha(t_n)$ remainder in (4.7) is $o(t_n^{3/2})$, proving the assertion about centering. □

The variance estimate also gives an almost sure growth rate. Here monotonicity of the occupied set allows a summable subsequence estimate to control all times.

Corollary 4.4. *Almost surely,*

$$\frac{L_n}{(\log n)^3} \longrightarrow \frac{2}{3\alpha^3}.$$

Proof. Let $n_k = \lfloor e^k \rfloor$. For every $\eta > 0$, (4.8) and Chebyshev's inequality give

$$\sum_{k \ge 1} \mathbb{P}(|L_{n_k} - \mu_{n_k}| > \eta k^3) < \infty,$$

since the summands are $O_{\alpha,\eta}(k^{-3})$. Borel–Cantelli, applied to $\eta = 1/m$, $m \in \mathbb{N}$, and (4.7) imply $L_{n_k}/(\log n_k)^3 \to 2/(3\alpha^3)$ almost surely. For $n_k \le n \le n_{k+1}$, monotonicity gives

$$\frac{L_{n_k}}{(\log n_{k+1})^3} \le \frac{L_n}{(\log n)^3} \le \frac{L_{n_{k+1}}}{(\log n_k)^3}.$$

Both outer terms have the asserted limit because $\log n_{k+1}/\log n_k \to 1$. □

References

[1] A. D. Barbour (2009). Univariate approximations in the infinite occupancy scheme. *ALEA* **6**, 415–433.

[2] A. D. Barbour and A. V. Gnedin (2009). Small counts in the infinite occupancy scheme. *Electronic Journal of Probability* **14**, 365–384. arXiv:0809.4387.

[3] A. D. Barbour and P. Hall (1984). On the rate of Poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society* **95**, 473–480.

[4] Z. Derbazi, A. Gnedin and A. Marynych (2024). Records in the Infinite Occupancy Scheme. *ALEA* **21**, 1475–1493. arXiv:2308.01739.

[5] E. Eriksson (2025; revised May 6, 2026). Edge Exchangeable Graphs: Connectedness, Gaussianity and Completeness. arXiv:2501.09511v2.

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

[7] A. Iksanov, Z. Kabluchko and V. Kotelnikova (2022; preprint 2021). A functional limit theorem for nested Karlin's occupancy scheme generated by discrete Weibull-like distributions. *Journal of Mathematical Analysis and Applications* **507**, article 125798. arXiv:2104.06948.

[8] S. Karlin (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics* **17**, 373–401.

[9] L. Le Cam (1960). An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics* **10**, 1181–1197.

[10] R. Michel (1987). An improved error bound for the compound Poisson approximation of a nearly homogeneous portfolio. *ASTIN Bulletin* **17**, 165–169.