

On the variance of the number of occupied boxes

Leonid V. Bogachev ^{a,1}, Alexander V. Gnedin ^b,
Yuri V. Yakubovich ^{b,*,2}

^a*Department of Statistics, University of Leeds, Leeds LS2 9JT, UK*

^b*Department of Mathematics, Utrecht University, P.O. Box 80010, 3508 TA
Utrecht, The Netherlands*

Abstract

We consider the occupancy problem where balls are thrown independently at infinitely many boxes with fixed positive frequencies. It is well known that the random number of boxes occupied by the first n balls is asymptotically normal if its variance V_n tends to infinity. In this work, we mainly focus on the opposite case where V_n is bounded, and derive a simple necessary and sufficient condition for convergence of V_n to a finite limit, thus settling a long-standing question raised by Karlin in the seminal paper of 1967. One striking consequence of our result is that the possible limit may only be a positive integer number. Some new conditions for other types of behavior of the variance, like boundedness or convergence to infinity, are also obtained. The proofs are based on the poissonization techniques.

Key words: Occupancy problem; Number of occupied boxes; Bounded variance; Poissonization; Geometric random variables

MSC: 60F05; 60C05

* Corresponding author.

Email addresses: bogachev@maths.leeds.ac.uk (Leonid V. Bogachev),
gnedin@math.uu.nl (Alexander V. Gnedin), yakubovich@math.uu.nl (Yuri V.
Yakubovich).

¹ Supported in part by DFG Grant 436 RUS 113/722 and a WUN GEP visiting grant.

² Supported by NWO Open Competition Grant 613.000.304.

1 Introduction

The classical occupancy problem is one of the cornerstones of discrete probability, dating back to its early ages (and hence encountered over and over again by the generations of students studying elementary probability through the evergreen hits like the birthday problem, the coupon collector's problem, etc. [1,15]). It still attracts lots of research interest, especially in recent years, mainly due to its numerous applications spreading across the board, from sampling statistics and quality control to quantum physics, bioinformatics and computer science. For an introduction to the field and a survey of the many models and results, see [10,21,24,27,28] and further references to original work therein.

In this paper, we are concerned with a version of the occupancy problem in an *infinite urn scheme* (first considered by Bahadur [3] and later on studied by Darling [11] and most systematically by Karlin [25]), in which the balls labeled $1, 2, \dots$ are thrown independently at an infinite array of boxes (*urns*) $j = 1, 2, \dots$, with fixed probability (*frequency*) p_j of hitting box j . The frequencies p_j are assumed to be strictly positive and satisfying

$$\|p\| := \sum_{j=1}^{\infty} p_j = 1. \quad (1.1)$$

Without loss of generality, we further assume that the sequence (p_j) is non-increasing, $p_1 \geq p_2 \geq \dots$.

Let K_n be the number of boxes discovered by the first n balls (i.e., occupied by at least one of the first n balls). Many other interpretations of this functional appear in the literature: for instance, when (p_j) is considered as a probability distribution on positive integers, K_n is the number of distinct values occurring among n random values sampled independently from (p_j) . Since there are infinitely many boxes, K_n increases unboundedly (with probability one) as more balls are thrown, which also implies (e.g., by Fatou's lemma) that the same is true for the expected number of occupied boxes, $\mathbb{E}(K_n)$. Moreover, as shown by Karlin [25, Theorem 8], $\lim_{n \rightarrow \infty} K_n / \mathbb{E}(K_n) = 1$ with probability one (an earlier result about convergence in probability was obtained by Bahadur [3]).

The more delicate asymptotic properties of the random variable K_n are largely determined by its variance $V_n := \text{Var}(K_n)$. It is known [13,20,25] that the distribution of K_n converges to a normal distribution provided that $V_n \rightarrow \infty$ as $n \rightarrow \infty$. The latter occurs, for instance, when the frequencies have a power-like decay, $p_j \sim cj^{-\alpha}$ ($j \rightarrow \infty$) with $\alpha > 1$ or, more generally, satisfy a condition of regular variation [25]. (Here and throughout, c stands for a generic positive constant, specific value of which is not important.)

1.1 Main result: the case of converging variance

In this paper, we essentially focus on the opposite situation, that is, when V_n is uniformly bounded (and hence the distribution of K_n does not converge to normal). In particular, we prove the following surprising characterization of frequencies (p_j) for which the variance V_n tends to a finite limit as $n \rightarrow \infty$.

Theorem 1.1. *A finite limit $v := \lim_{n \rightarrow \infty} V_n$ exists if and only if for some integer $k \geq 1$ the frequencies satisfy the “lagged ratio” condition*

$$\lim_{j \rightarrow \infty} \frac{p_{j+k}}{p_j} = \frac{1}{2}, \quad (1.2)$$

and in this case the limiting value v coincides with the lag k .

The striking consequence of this result is that whenever the finite limit of the sequence (V_n) exists, it must be a positive integer number, $v \in \mathbb{N}$.

The issue of converging variance was first queried in the seminal paper by Karlin [25], where in particular he appreciated as “formidable if not impossible” the task to determine the behavior of the variance V_n without some regularity assumptions. In particular, adopting the condition of regular variation of the frequency tail, he came up with a sufficient condition for the existence of a finite limit of V_n [25, Theorem 2]. In fact, as we shall see below (in Section 5), convergence to a finite limit, combined with the special dyadic structure of the counting measure controlling the frequency input, is a regularity condition in itself, being strong enough to ensure the result of Theorem 1.1. (To be more precise, the “dyadic” feature mentioned above, pertains primarily to the *poissonized* version of the problem, i.e., with randomized number of balls, see Section 2 below.)

The prototypical (apparently folklore) instance of frequencies (p_j) with converging variance V_n is the geometric sequence of ratio $1/2$ (i.e., $p_j = 2^{-j}$), where one can show with some effort that $V_n \rightarrow 1$ as $n \rightarrow \infty$ (see [13,20,25]). Note that our condition (1.2) is obviously satisfied here with $k = 1$, hence the result. The mechanism leading to such a simple answer is due to a resonance of the ratio $q = 1/2$ with the intrinsic dyadic structure of the variance, resulting in massive cancelation of oscillating terms (again, in the poissonized version, see Example 2.3 below). Recently, such cancelations have been explained directly for the original model (i.e., for V_n) using sophisticated analytic methods [2,31].

It seems to be less well known that for generic geometric frequencies $p_j = cq^{-j}$, the (finite) limit of V_n exists if $q = 2^{-1/k}$ ($k \in \mathbb{N}$), with the limiting value $v = k$ (see [23, §4, page 15]). Again, using Theorem 1.1 one gets this

answer immediately, together with the “only if” statement; moreover, the same conclusion can be readily extended to sequences (p_j) from the parametric class RT_q (see [6,9,18]), defined by the property

$$\lim_{j \rightarrow \infty} \frac{p_{j+1}}{p_j} = q, \quad (1.3)$$

thus asymptotically mimicking the geometric decay. (Some concrete examples of distributions in the RT_q class, complementing the geometric instance, will be given below in Section 1.3.) Indeed, in the RT_q case equation (1.2) amounts to $q^k = 1/2$, whence $q = 2^{-1/k}$. Of course, condition (1.3) is too restrictive for the criterion (1.2), as can be seen for instance by merging k geometric sequences of the same ratio $q = 1/2$ (and normalizing the resulting sequence so as to satisfy (1.1)).

The following “decomposition” interpretation of Theorem 1.1 clarifies the compound structure of frequency sequences (p_j) that exhibit convergence of the variance. Observe that by condition (1.2), the sequence (p_j) splits in a disjoint fashion into k non-increasing subsequences $p_j^{(i)} := p_{i+k(j-1)}$ ($i = 1, \dots, k$), each belonging to the $\text{RT}_{1/2}$ class:

$$(p_j) = \bigsqcup_{i=1}^k (p_j^{(i)}) : \quad \lim_{j \rightarrow \infty} \frac{p_{j+1}^{(i)}}{p_j^{(i)}} = \frac{1}{2} \quad (i = 1, \dots, k). \quad (1.4)$$

Moreover, by the “if” part of Theorem 1.1, each of the k constituent subsequences brings a unit contribution to the overall limiting variance $v = k$.

Such a decomposition may be interpreted as splitting the initial array of boxes $1, 2, \dots$ into k infinite sub-arrays $\{i+k(j-1), j = 1, 2, \dots\}$ ($i = 1, \dots, k$), and allocating the balls to boxes in a two-stage procedure as follows: for each ball, a destination array is chosen independently with probabilities $\|p^{(i)}\|$, and the ball is then thrown with the corresponding (re-scaled) frequencies $p_j^{(i)} / \|p^{(i)}\|$ ($j = 1, 2, \dots$). The additivity of the variance in this procedure, as predicted by Theorem 1.1, may be somewhat surprising, given the apparent dependence of the partial occupancy numbers $K_n^{(i)}$ ($i = 1, \dots, k$). However, additivity becomes quite transparent in the poissonized setting, where the dependence between boxes is removed (see Remark 2.4 below).

1.2 Geometric frequencies

Historically, there has been some confusion about the converging variance in the geometric model. Controversy started in [25, Example 6], where Karlin asserted that his sufficient condition for convergence [25, Theorem 2] was satisfied for *every* geometric sequence $p_j = cq^j$ ($0 < q < 1$), with the limiting

value given by $v = \log_{1/q} 2$. As we have seen, this is false unless q belongs to the countable set $\{2^{-1/k}, k \in \mathbb{N}\}$. A more careful inspection reveals that Karlin's condition, if applied accurately, does yield the correct answer in the geometric case, properly discriminating between convergence vs. divergence! Moreover, we have found out, quite unexpectedly, that Karlin's condition (decorated in [25] with some superfluous assumptions and originally conceived as just a sufficient condition) proves to be *necessary and sufficient*, being equivalent to our own criterion proved in Lemma 5.1. We will discuss this link below, in Section 5.4.

That there was something wrong with Example 6 in [25] was subsequently pointed out by Dutko [13, page 1258], who noticed that V_n is bounded below by a positive constant, uniformly in n and q , hence the limit $v = \log_{1/q} 2$ cannot be valid at least for small values of q (when $\log_{1/q} 2$ gets arbitrarily close to zero). However, Dutko [13, page 1258] apparently claimed that the limit of the variance fails to exist for *each* $q \neq 1/2$, thus missing the other values, $q = 2^{-1/k}$, $k > 1$. Unfortunately, he gave no details to support such a conclusion, referring to his unpublished thesis [12], which is not easily available.

More recent studies [2,19,29,31] have shed much light on the geometric model. Hitczenko and Louchard [19] (motivated by random compositions of natural numbers) were apparently first to prove analytically that $V_n = 1 + o(1)$ in the geometric case with $q = 1/2$, contrary to "popular belief" [31] that persistent oscillations are ubiquitous in discrete random structures involving geometric distribution (see, e.g., [20,32,33]). Prodinger [31] gave an alternative proof of this asymptotics (along with a similar result for a particular model of data search trees called PATRICIA tries), proceeding from the general "oscillatory" framework. Recently, Archibald *et al.* [2, Theorem 2] derived a very precise asymptotic expansion

$$V_n = \log_{1/q} 2 + \delta_V(\log_{1/q} n) + o(1) \quad (n \rightarrow \infty), \quad (1.5)$$

where $\delta_V(x) := \delta_E(x + \log_{1/q} 2) - \delta_E(x)$ with $\delta_E(\cdot)$ periodic of period 1 and zero mean (the latter function emerges in a similar expansion for Φ_n , the expected value of K_n). If $q = 1/2$ then $\log_{1/q} 2 = 1$, and from the expansion (1.5) it is seen that the oscillating term vanishes due to 1-periodicity of $\delta_E(\cdot)$, since $\delta_V(x) = \delta_E(x + 1) - \delta_E(x) = 0$ (see [2, Appendix A, page 1079]). In fact, the same argument is true for any $q = 2^{-1/k}$ ($k \in \mathbb{N}$), when $\log_{1/q} 2 = k$ and hence $\delta_V(x) = \delta_E(x + k) - \delta_E(x) = 0$ (see [23, §4, page 15]).

1.3 Bounded variance and convergence to infinity

One can also wonder about conditions for other possible types of behavior of the variance V_n . We shall prove the following criterion of uniform boundedness,

again set in terms of the lagged ratio p_{j+k}/p_j compared to the *upper* threshold $1/2$ (cf. (1.2)).

Theorem 1.2. *The sequence (V_n) is bounded if and only if there exists a positive integer k such that the frequencies (p_j) satisfy the condition*

$$\limsup_{j \rightarrow \infty} \frac{p_{j+k}}{p_j} \leq \frac{1}{2}. \quad (1.6)$$

Moreover, if k is the least integer with the property (1.6), then (V_n) satisfies a sharp asymptotic bound $\limsup_{n \rightarrow \infty} V_n \leq k$.

This situation is exemplified by the generic geometric frequencies, with arbitrary ratio $0 < q < 1$. Another example is the Poisson frequencies $p_j = c\lambda^j/j!$ ($\lambda > 0$), where the variance V_n is bounded but does not converge: indeed, here $p_{j+k}/p_j \sim (\lambda/j)^k \rightarrow 0$ as $j \rightarrow \infty$, hence (1.6) is fulfilled whereas (1.2) fails. A larger class is that of quasi-binomial distributions [26], given by $p_j = (c/j!) \prod_{i=0}^{j-1} (\lambda + iq)$ with parameters $\lambda > 0$, $0 \leq q < 1$. (To explain the name, note that $c^{-1} = (1 - q)^{-\lambda/q} - 1$ for $q > 0$, while for $q = 0$ one has, in a continuous fashion, $c^{-1} = e^\lambda - 1$, thus recovering the Poisson normalization constant.) Somewhat similar but different parametric family is given by the negative binomial distribution $p_j = (cq^j/j!) \prod_{i=0}^{j-1} (\lambda + i) = c \binom{\lambda+j-1}{j} q^j$, with $\lambda > 0$, $0 < q < 1$ [here $c^{-1} = (1 - q)^{-\lambda} - 1$].

Note that all these examples belong to classes RT_q with $0 \leq q < 1$. It is possible to construct more general examples using the “decomposition” reformulation of Theorem 1.2 in the spirit of (1.4), in that the variance V_n is uniformly bounded if and only if the sequence (p_j) may be split in a disjoint fashion into a finite number of subsequences, each of which satisfies condition (1.6) with $k = 1$ (e.g., each from RT_{q_i} with $0 \leq q_i \leq 1/2$, $i = 1, \dots, k$).

We shall also address the classical question of convergence to infinity and produce new conditions ensuring that $V_n \rightarrow \infty$. Note, however, that in contrast to the convergent or bounded cases, no necessary and sufficient criteria are available without extra regularity assumptions. To illustrate our results in this direction, let us formulate here two sufficient conditions, the first of which is set in terms of the lagged ratios p_{j+k}/p_j against the *lower* threshold $1/2$ (cf. (1.6)), while the second one is based on the “tail ratio”

$$\rho_j := \frac{1}{p_j} \sum_{i>j} p_i. \quad (1.7)$$

Theorem 1.3. *Suppose that for each integer $k \geq 1$,*

$$\liminf_{j \rightarrow \infty} \frac{p_{j+k}}{p_j} \geq \frac{1}{2}. \quad (1.8)$$

Then it follows that

$$\lim_{j \rightarrow \infty} \rho_j = \infty, \quad (1.9)$$

which in turn implies that $V_n \rightarrow \infty$ as $n \rightarrow \infty$.

Examples to Theorem 1.3 are immediately supplied by the class RT_1 , where condition (1.8) is obviously satisfied for any $k \geq 1$. More complex examples (not in RT_1) will be constructed in Sections 4.1 and 4.3.

Remark 1.4. The tail ratio (1.7) can be expressed as $\rho_j = (1 - h_j)/h_j$, where $h_j = p_j / \sum_{i \geq j} p_i$ is the discrete-time hazard rate, a key characteristic in reliability theory and survival analysis (see, e.g., [4]). The latter quantity also appears in the extreme value theory in connection with records from discrete distributions, where it is interpreted as the probability that j is a record value (see, e.g., [30,34]). In the occupancy context, condition (1.9) is related to the ‘‘probability of a tie for first place’’ $\mathbb{P}\{X_{n,M_n} = 1\}$, where $M_n := \max\{j : X_{n,j} \neq 0\}$ is the largest index among the occupied boxes after n throws. More specifically, it has been proved [5,14] that condition (1.9) is satisfied if and only if

$$\mathbb{P}\{X_{n,M_n} = 1\} \rightarrow 1 \quad (n \rightarrow \infty), \quad (1.10)$$

and moreover, if (1.9) fails then $\mathbb{P}\{X_{n,M_n} = 1\}$ does not converge at all. This, combined with Theorem 4.3, shows that (1.10) implies both $V_n \rightarrow \infty$ and $\Phi_{n,1} \rightarrow \infty$, which is a surprising connection between the behavior in the extreme-value range and the global characteristics of the sample. These facts equally apply to the poissonized model.

1.4 Outline

The rest of the paper is organized as follows. Section 2 contains general formulas and introduces the poissonization technique. In Section 3, we connect the variance V_n with the mean number of singletons (i.e., the boxes occupied by exactly one of the first n balls) and derive useful upper bounds. We also obtain here a basic integral representation of the poissonized variance $V(t)$ via the Laplace transform of the function $\Delta\nu(x)$, counting the frequencies p_j in the interval $[x/2, x]$, and relate the threshold values of $\Delta\nu(\cdot)$ with the lagged ratios p_{j+k}/p_j . This analysis culminates in the proof of Theorem 1.2. In Section 4, various sufficient conditions for $V_n \rightarrow \infty$ are derived, which covers the content of Theorem 1.3. We also show that these conditions are not necessary, by constructing examples of weird oscillatory behavior. In Section 5, we derive a simple integral condition in terms of the function $\Delta\nu(\cdot)$, necessary and sufficient in order that $V(t)$ converge to a finite limit. This criterion is then used to prove Theorem 1.1. In conclusion, we rehabilitate Karlin’s sufficient condition of convergence, by showing that it is in fact necessary and sufficient.

2 Poissonization and moment formulas

Let $X_{n,j}$ be the occupancy number of box j after n throws, that is, the number of balls out of the first n that land in box j . Note that

$$K_n = \sum_{j=1}^{\infty} \mathbf{1}\{X_{n,j} > 0\}, \quad (2.1)$$

where $\mathbf{1}(A)$ is the indicator of event A (i.e., with values 1 when A is true and 0 otherwise). Because $\sum_{j=1}^{\infty} X_{n,j} = n$, it is clear that the terms in the sum (2.1) are not independent.

2.1 Poissonization

A common recipe to circumvent the dependence (see [1,22] for a general introduction and [20,24,25,27] for details in the occupancy problem context) is to consider a closely related model in which the balls are thrown at the jump times of a unit rate Poisson process ($N(t)$, $t \geq 0$): by this randomization the balls appear in boxes according to independent Poisson processes $X_j(t)$, with rate p_j for box j . Further advantage of the poissonized model is that the normalization (1.1) can be replaced by a weaker summability condition $\|p\| \equiv \sum_{j=1}^{\infty} p_j < \infty$, thus allowing one to avoid computing normalization constants in expressions for p_j . Clearly, the normalization (1.1) can always be maintained by rescaling the frequencies $p_j \mapsto \|p\|^{-1}p_j$, to the effect of a linear time change, $t \mapsto \|p\|t$.

In what follows, we adopt the convention that quantities derived from the poissonized version of the occupancy problem are written as functions of the continuous time parameter t , while for the original model we preserve the notation with lower index n . In particular, we write $X_j(t)$ (cf. above) for the number of balls that land in box j by time t and

$$K(t) := K_{N(t)} = \sum_{j=1}^{\infty} \mathbf{1}\{X_j(t) > 0\} \quad (2.2)$$

for the number of boxes discovered by the Poisson process $N(t)$. Likewise, denoting by $K_{n,r}$ the number of boxes, each of which is hit by exactly r of the first n balls, we write

$$K_r(t) := K_{N(t),r} = \sum_{j=1}^{\infty} \mathbf{1}\{X_j(t) = r\}$$

for the corresponding poissonized quantity (which is the number of boxes

containing exactly r balls each by time t). Clearly,

$$\begin{aligned} K_n &= \sum_r K_{n,r}, & K(t) &= \sum_r K_r(t), \\ n &= \sum_r r K_{n,r}, & N(t) &= \sum_r r K_r(t). \end{aligned} \quad (2.3)$$

For the mean values of the number of occupied boxes we have the formulas

$$\Phi_n := \mathbb{E}(K_n) = \sum_{j=1}^{\infty} (1 - (1 - p_j)^n), \quad (2.4)$$

$$\Phi(t) := \mathbb{E}(K(t)) = \sum_{j=1}^{\infty} (1 - e^{-tp_j}), \quad (2.5)$$

related by the poissonization identity

$$\Phi(t) = e^{-t} \sum_{n=0}^{\infty} \frac{t^n}{n!} \Phi_n,$$

where $\Phi_0 = 0$. Encoding the collection of frequencies into an infinite counting measure on $\mathbb{R}_+ =]0, \infty[$

$$\nu(dx) := \sum_{j=1}^{\infty} \delta_{p_j}(dx) \quad (2.6)$$

(where δ_x is the Dirac mass at x , i.e., $\delta_x(A) = \mathbf{1}\{x \in A\}$ for $A \subset \mathbb{R}_+$), we can represent the mean values (2.4), (2.5) in an integral form as

$$\Phi_n = \int_0^1 (1 - (1 - x)^n) \nu(dx), \quad (2.7)$$

$$\Phi(t) = \int_0^{\infty} (1 - e^{-tx}) \nu(dx). \quad (2.8)$$

Remark 2.1. When the frequencies are normalized by (1.1) then all $p_j \leq 1$ and the integral in (2.8) could be written in the limits from 0 to 1, similarly to (2.7). In the poissonized model, specific normalization is not important, so we prefer to use a more flexible notation as in (2.8). The same convention applies to similar representations below (see, e.g., formulas (2.10) and (2.14)).

Furthermore, set

$$\Phi_{n,r} := \mathbb{E}(K_{n,r}) = \binom{n}{r} \int_0^1 x^r (1 - x)^{n-r} \nu(dx), \quad (2.9)$$

$$\Phi_r(t) := \mathbb{E}[K_r(t)] = \frac{t^r}{r!} \int_0^{\infty} x^r e^{-tx} \nu(dx), \quad (2.10)$$

the latter being related to the derivatives of $\Phi(t)$ via

$$\Phi_r(t) = (-1)^{r+1} \frac{t^r}{r!} \Phi^{(r)}(t). \quad (2.11)$$

Note that equations (2.3) imply

$$\begin{aligned}\Phi_n &= \sum_r \Phi_{n,r}, & \Phi(t) &= \sum_r \Phi_r(t), \\ n &= \sum_r r \Phi_{n,r}, & t &= \sum_r r \Phi_r(t).\end{aligned}\tag{2.12}$$

An analyst will recognize in (2.8) a Bernstein function (see [7]) with the following general properties (see also [17]).

Lemma 2.2. *If an infinite measure ν on \mathbb{R}_+ satisfies $\int_0^\infty (1 - e^{-x}) \nu(dx) < \infty$, then (2.8) defines a function $\Phi(\cdot)$ which*

- (i) *is analytic in the right half-plane,*
- (ii) *has alternating derivatives $(-1)^{r+1} \Phi^{(r)}(t) > 0$ ($t > 0$),*
- (iii) *satisfies $\Phi(t) \uparrow \infty$ but $\Phi(t) = o(t)$ as $t \rightarrow \infty$.*

Conversely, if a function $\Phi(t)$ on $[0, \infty[$ has the properties (ii) and (iii) along with $\Phi(0) = 0$, then there exists a unique infinite measure ν on \mathbb{R}_+ such that representation (2.8) holds.

2.2 The variance of the number of occupied boxes

By the independence of summands in (2.2), the variance of $K(t)$ is given by

$$V(t) := \text{Var}(K(t)) = \sum_{j=1}^{\infty} (e^{-tp_j} - e^{-2tp_j}),\tag{2.13}$$

which is the same as

$$V(t) = \int_0^\infty (e^{-tx} - e^{-2tx}) \nu(dx) = \Phi(2t) - \Phi(t).\tag{2.14}$$

Example 2.3. For geometric frequencies of ratio $q = 1/2$, that is, $p_j = 2^{-j}$ ($j = 1, 2, \dots$), the sum (2.13) is evaluated explicitly thanks to telescoping of partial sums (see [13, page 1258]):

$$V(t) = \lim_{M \rightarrow \infty} \sum_{j=1}^M (e^{-t2^{-j}} - e^{-t2^{-j+1}}) = \lim_{M \rightarrow \infty} (e^{-t2^{-M}} - e^{-t}) = 1 - e^{-t}.$$

In particular, it follows that $V(t) \rightarrow 1$ as $t \rightarrow \infty$. More generally, a similar simplification occurs in the geometric case with the ratio $q = 2^{-1/k}$ ($k \geq 1$), where it is convenient to split the sum in (2.13) into k sub-sums (over $j = i + k(\ell - 1)$, where $i = 1, \dots, k$, $\ell = 1, 2, \dots$), each involving a (non-normalized) geometric sequence with ratio $1/2$. Applying the previous result (with $q = 1/2$)

and adding up the k unit contributions emerging in the limit from the k constituent subsequences, we obtain the convergence $V(t) \rightarrow k$ as $t \rightarrow \infty$. For other values of q the formula for the variance does not simplify.

Remark 2.4. The poissonized variance is additive: if $(p_j^{(1)})$ and $(p_j^{(2)})$ are two summable sequences of frequencies, and if (p_j) is obtained by merging them into a single sequence, then the corresponding variances satisfy $V^{(1)}(t) + V^{(2)}(t) = V(t)$. This explains the structural decomposition of the variance mentioned in the Introduction and illustrated in Example 2.3.

The fixed- n counterpart of (2.13) is

$$V_n = \Phi_{2n} - \Phi_n + \sum_{i \neq j}^{\infty} \left((1 - p_i - p_j)^n - (1 - p_i)^n (1 - p_j)^n \right), \quad (2.15)$$

where the cross-terms arise due to dependence in (2.1).

2.3 Depoissonization

According to [20, Proposition 4.3(ii)], the variances $V(n)$ and V_n are always of the same order,

$$0 < \liminf_{n \rightarrow \infty} \frac{V(n)}{V_n} \leq \limsup_{n \rightarrow \infty} \frac{V(n)}{V_n} < \infty. \quad (2.16)$$

In the next lemma, we establish estimates for the deviation of the poissonized quantities from their fixed- n counterpart in terms of higher-order moments, which will be instrumental for depoissonization in the case of bounded variance (see Section 3).

Lemma 2.5. *If the normalization (1.1) holds then*

$$\Phi(n) - \Phi_n = O(n^{-1}) \Phi_2(n), \quad (2.17)$$

$$V(n) - V_n = O(n^{-1}) \left(\Phi_1(n)^2 + \Phi_2(n) \right), \quad (2.18)$$

and for each $r = 1, 2, \dots$

$$\Phi_r(n) - \Phi_{n,r} = O(n^{-1}) \left(\Phi_r(n) + \Phi_{r+1}(n) + \Phi_{r+2}(n) \right), \quad (2.19)$$

Proof. We shall need the elementary inequalities

$$0 \leq e^{-nx} - (1 - x)^n \leq nx^2 e^{-nx} \quad (0 \leq x \leq 1). \quad (2.20)$$

The first inequality is obvious, while the second one follows from the estimate

$$(1-x)^n \geq (1-x^2)^n e^{-nx} \geq (1-nx^2) e^{-nx}.$$

Now, using representations (2.7), (2.8) (rewriting the integral (2.8) in the limits from 0 to 1, due to (1.1)) and inserting the bounds (2.20), we obtain

$$0 \leq \Phi_n - \Phi(n) = \int_0^1 (e^{-nx} - (1-x)^n) \nu(dx) \leq \frac{2}{n} \Phi_2(n),$$

which proves (2.17). Next, from (2.9) and (2.10) we get

$$\Phi_r(n) - \Phi_{n,r} = O(n^{-1}) \Phi_r(n) + \frac{n^r}{r!} \int_0^1 x^r (e^{-nx} - (1-x)^{n-r}) \nu(dx). \quad (2.21)$$

By the inequalities (2.20), for each $x \in [0, 1]$

$$e^{-nx} - (1-x)^{n-r} \geq e^{-nx} - e^{-(n-r)x} \geq -(e^r - 1)x e^{-nx}, \quad (2.22)$$

$$e^{-nx} - (1-x)^{n-r} \leq e^{-nx} - (1-x)^n \leq nx^2 e^{-nx}. \quad (2.23)$$

Substituting the estimates (2.22) and (2.23) into (2.21) and recalling the notation (2.10) yields (2.19).

Finally, as shown in [20, Theorem 2.3], the cross-terms in (2.15) can be evaluated as

$$(1-p_i)^n (1-p_j)^n - (1-p_i-p_j)^n = np_i p_j (1-p_i)^{n-1} (1-p_j)^{n-1} + O(n^2 p_i^2 p_j^2 (1-p_i)^{n-2} (1-p_j)^{n-2}).$$

Inserting this estimate into (2.15) and summing over all i, j , we obtain

$$V_n = \Phi_{2n} - \Phi_n + O(n^{-1}) \Phi_{n,1}^2 + O(n^{-2}) \Phi_{n,2}^2. \quad (2.24)$$

From (2.12) and (2.7) it follows that if the condition (1.1) holds then

$$\Phi_{n,r} \leq \Phi_n = \int_0^1 (1 - (1-x)^n) \nu(dx) \leq \int_0^1 nx \nu(dx) = n,$$

and similarly, using (2.8),

$$\Phi_r(n) \leq \Phi(n) = \int_0^\infty (1 - e^{-nx}) \nu(dx) \leq \int_0^\infty nx \nu(dx) = n.$$

Hence, subtracting (2.24) from (2.14) and using the estimates (2.17) and (2.19), we arrive at (2.18). \square

3 Bounded variance

In this section, we mainly focus on the situation where the variance $V(t)$ is bounded.

3.1 Auxiliary estimates

We first derive various useful inequalities involving the functions $V(t)$, $\Phi(t)$, $\Phi_r(t)$ and the measure ν . Since $\Phi'(t)$ is decreasing and $V(t) = \Phi(2t) - \Phi(t)$, the mean value theorem yields

$$\Phi'(2t) \leq \frac{\Phi(2t) - \Phi(t)}{t} = \frac{V(t)}{t} \leq \Phi'(t),$$

or equivalently

$$\frac{1}{2}\Phi_1(2t) \leq V(t) \leq \Phi_1(t). \quad (3.1)$$

The first inequality in (3.1) generalizes.

Lemma 3.1. *For $r = 1, 2, \dots$ and $t > 0$,*

$$\Phi_r(t) \leq \frac{2^{r(r+1)/2}}{r!} V(2^{-r}t).$$

Proof. Recalling the definition (2.11) and setting $f_r(t) := (-1)^{r+1}\Phi^{(r)}(t) > 0$ (see Lemma 2.2(ii)), we shall prove by induction the equivalent inequality

$$f_r(t) \leq \frac{2^{r(r+1)/2} V(2^{-r}t)}{t^r} \quad (t > 0). \quad (3.2)$$

For $r = 1$ inequality (3.2) follows from the first inequality in (3.1). Suppose (3.2) holds for f_1, \dots, f_{r-1} . Note that $f_{r-1}''(t) = f_{r+1}(t) > 0$, hence the function f_{r-1} is convex and therefore

$$\frac{f_{r-1}(t/2) - f_{r-1}(t)}{t/2} \geq -f_{r-1}'(t) = f_r(t). \quad (3.3)$$

On the other hand, since $f_{r-1}(t) \geq 0$ and by the induction hypothesis,

$$\frac{f_{r-1}(t/2) - f_{r-1}(t)}{t/2} \leq \frac{f_{r-1}(t/2)}{t/2} \leq \frac{2^{r(r-1)/2} V(2^{-r}t)}{(t/2)^r}. \quad (3.4)$$

Combining (3.3) and (3.4), we obtain (3.2) for f_r . Thus, the induction step follows, and the proof is complete. \square

Consider the limits superior

$$\bar{v} := \limsup_{t \rightarrow \infty} V(t), \quad \bar{\varphi}_r := \limsup_{t \rightarrow \infty} \Phi_r(t) \quad (r = 1, 2, \dots). \quad (3.5)$$

By continuity, $V(t)$ is uniformly bounded on $[0, \infty[$ if and only if $\bar{v} < \infty$, and the same is true for $\Phi_r(t)$ in terms of the condition $\bar{\varphi}_r < \infty$.

Note that \bar{v} is strictly positive (cf. [13, page 1258]); indeed, setting $t = 1/p_k$ in (2.13) we have

$$\bar{v} \geq \limsup_{k \rightarrow \infty} \sum_{j=1}^{\infty} \left(e^{-p_j/p_k} - e^{-2p_j/p_k} \right) \geq e^{-1} - e^{-2} > 0. \quad (3.6)$$

Corollary 3.2. *The conditions $\bar{v} < \infty$ and $\bar{\varphi}_1 < \infty$ are equivalent and imply $\bar{\varphi}_r < \infty$ for all $r \geq 1$.*

Proof. Follows from (3.1) and Lemma 3.1. □

Appealing to Lemma 2.5, we have depoissonization in terms of moments.

Corollary 3.3. *If $\bar{v} < \infty$ then, as $n \rightarrow \infty$,*

$$\Phi(n) - \Phi_n = O(n^{-1}), \quad V(n) - V_n = O(n^{-1}),$$

and, for all $r \geq 1$,

$$\Phi_r(n) - \Phi_{n,r} = O(n^{-1}).$$

3.2 Uniform upper bounds for $\bar{\varphi}_r$

Lemma 3.1 entails an estimate of $\bar{\varphi}_r$ through either \bar{v} or $\bar{\varphi}_1$. With some more effort, we will derive an improved upper bound that does not depend on r . Recall that the measure ν is defined in (2.6), and consider the new (finite) measure

$$\tilde{\nu}(\mathrm{d}x) := x\nu(\mathrm{d}x) = \sum_{j=1}^{\infty} p_j \delta_{p_j}(\mathrm{d}x). \quad (3.7)$$

When the normalization (1.1) holds, this is a probability measure governing the frequency distribution of the random box discovered by ball 1.

Using the measure $\tilde{\nu}$, we can rewrite (2.10) as follows

$$\Phi_r(t) = \frac{t^r}{r!} \int_0^{\infty} x^{r-1} e^{-xt} \tilde{\nu}(\mathrm{d}x). \quad (3.8)$$

Also, let us set

$$\bar{\eta} := \limsup_{x \downarrow 0} \frac{\tilde{\nu}[0, x]}{x}. \quad (3.9)$$

Lemma 3.4. *Suppose that $\bar{v} < \infty$. Then for all $r = 1, 2, \dots$*

$$\bar{\varphi}_r \leq \bar{\eta} \leq e\bar{\varphi}_1 \leq 2e\bar{v}. \quad (3.10)$$

Proof. Note that the last inequality in (3.10) follows from (3.1). Further, integrating by parts in (3.8) and using the substitution $y = xt$, we get

$$\Phi_r(t) = \frac{t}{r!} \int_0^\infty e^{-y} y^{r-2} (y + 1 - r) \tilde{\nu}[0, y/t] dy. \quad (3.11)$$

For $r = 1$, due to monotonicity of the function $\tilde{\nu}[0, \cdot]$, (3.11) implies

$$\Phi_1(t) \geq t \int_1^\infty e^{-y} \tilde{\nu}[0, y/t] dy \geq e^{-1} \frac{\tilde{\nu}[0, 1/t]}{1/t}, \quad (3.12)$$

and by letting here $t \rightarrow \infty$ we obtain $\bar{\varphi}_1 \geq e^{-1} \bar{\eta}$ (see (3.9), (3.10)).

On the other hand, for any $r \geq 1$ from (3.11) it follows that

$$\Phi_r(t) \leq \frac{1}{r!} \int_0^\infty e^{-y} y^r \frac{\tilde{\nu}[0, y/t]}{y/t} dy \quad (r \geq 1),$$

which implies $\bar{\varphi}_r \leq \bar{\eta}$ by the ‘‘lim sup’’ part of Fatou’s lemma [16, §IV.2]. \square

3.3 Growth of the mean number of occupied boxes

Lemma 3.4 implies that if $\bar{v} < \infty$ then each term in the decomposition $\Phi(t) = \sum_{r=1}^\infty \Phi_r(t)$ makes a uniformly bounded contribution to $\Phi(t) \rightarrow \infty$. This is to be contrasted with the case of frequencies akin to $p_j \sim cj^{-\alpha}$ ($\alpha > 1$), where $V(t)$, $\Phi(t)$ and $\Phi_r(t)$ ($r \geq 1$) are of the same order $O(t^{1/\alpha})$ as $t \rightarrow \infty$ (see [25]). The next lemma estimates the growth of $\Phi(t)$ in the case of bounded variance.

Lemma 3.5. *Suppose that $\bar{v} < \infty$. Then*

$$\limsup_{t \rightarrow \infty} \frac{\Phi(t)}{\log t} \leq 2\bar{v}.$$

Proof. For any $\varepsilon > 0$, there exists $t_0 > 0$ such that for all $t \geq t_0$

$$\Phi_1(t) \leq \bar{\varphi}_1 + \varepsilon \leq 2\bar{v} + \varepsilon,$$

due to Lemma 3.4. Therefore,

$$\Phi(t) - \Phi(t_0) = \int_{t_0}^t \Phi'(s) \, ds = \int_{t_0}^t \frac{\Phi_1(s)}{s} \, ds \leq (2\bar{\nu} + \varepsilon)(\log t - \log t_0).$$

Hence,

$$\limsup_{t \rightarrow \infty} \frac{\Phi(t)}{\log t} = \limsup_{t \rightarrow \infty} \frac{\Phi(t) - \Phi(t_0)}{\log t - \log t_0} \leq 2\bar{\nu} + \varepsilon,$$

and since $\varepsilon > 0$ is arbitrary, our claim follows.

A shorter proof is by a simple “lim sup” version of L’Hôpital’s rule:

$$\limsup_{t \rightarrow \infty} \frac{\Phi(t)}{\log t} \leq \limsup_{t \rightarrow \infty} \frac{\Phi'(t)}{1/t} = \limsup_{t \rightarrow \infty} \Phi_1(t) = \bar{\varphi}_1 \leq 2\bar{\nu},$$

due to Lemma 3.4. □

3.4 The basic representation of the variance $V(t)$

As in [25], it is convenient to rewrite the formula (2.14) for the variance as a single integral representation. Recall that ν is given by (2.6), and introduce the function

$$\Delta\nu(x) := \nu]x/2, x] = \#\{j : x/2 < p_j \leq x\} \quad (x > 0). \quad (3.13)$$

Lemma 3.6. *The variance $V(t)$ can be represented as*

$$V(t) = t \int_0^\infty e^{-tx} \Delta\nu(x) \, dx \quad (t \geq 0). \quad (3.14)$$

Proof. Let us rewrite the definition (3.13) as

$$\Delta\nu(x) = \sum_{j=1}^{\infty} \mathbf{1}\{p_j \leq x < 2p_j\}.$$

Substituting this representation into the right-hand side of equation (3.14) and interchanging the order of summation and integration, we obtain

$$\begin{aligned} t \int_0^\infty e^{-tx} \Delta\nu(x) \, dx &= t \int_0^\infty e^{-tx} \sum_{j=1}^{\infty} \mathbf{1}\{p_j \leq x < 2p_j\} \, dx \\ &= t \sum_{j=1}^{\infty} \int_{p_j}^{2p_j} e^{-tx} \, dx = \sum_{j=1}^{\infty} (e^{-tp_j} - e^{-2tp_j}) = V(t), \end{aligned}$$

by equation (2.13). □

Corollary 3.7. *The function*

$$D(x) := \int_0^x \Delta\nu(u) \, du \quad (3.15)$$

is well defined and uniformly bounded for all $x \geq 0$. In particular, $D(0) = 0$.

Proof. Letting $t = 1$ in (3.14), we obtain

$$V(1) \geq \int_0^x e^{-u} \Delta\nu(u) \, du \geq e^{-x} \int_0^x \Delta\nu(u) \, du,$$

hence $D(x) \leq e^x V(1) < \infty$ for any $x > 0$. Vanishing at zero is obtained by the absolute continuity of the integral. Finally, boundedness of $D(x)$ follows because $\Delta\nu(x) \equiv 0$ for all x large enough. \square

Integrating by parts in (3.14) and using Corollary 3.7, we obtain an alternative representation, which will also be useful:

$$V(t) = t^2 \int_0^\infty e^{-tx} D(x) \, dx = \int_0^\infty e^{-y} y \frac{D(y/t)}{y/t} \, dy \quad (t > 0). \quad (3.16)$$

3.5 Estimates using the function $\Delta\nu(x)$

It is immediately clear from (3.14) that if $\Delta\nu(x) \leq c$ for all $x > 0$ then $V(t) \leq c$ for all $t > 0$. Moreover, one can obtain two-sided asymptotic bounds as follows.

Lemma 3.8. *Recall that \bar{v} is given by (3.5), and set*

$$\bar{w} := \limsup_{x \downarrow 0} \Delta\nu(x).$$

Then $\bar{v} < \infty$ if and only if $\bar{w} < \infty$, and in this case

$$(\sqrt{5} - 2) \bar{w} \leq \bar{v} \leq \bar{w}. \quad (3.17)$$

Proof. The substitution $y = tx$ in (3.14) yields

$$V(t) = \int_0^\infty e^{-y} \Delta\nu(y/t) \, dy,$$

and an application of the “lim sup” part of Fatou’s lemma [16, §IV.2] implies

$$\bar{v} \leq \bar{w} \int_0^\infty e^{-y} \, dy = \bar{w}.$$

For the converse inequality, we need to exploit the special structure of the measure ν . Fixing $x > 0$ and retaining in (2.13) the terms with $p_j \in]x/2, x]$ only, we obtain

$$V(t) \geq \Delta\nu(x) \min_{p \in [x/2, x]} (e^{-tp} - e^{-2tp}). \quad (3.18)$$

It is clear that the minimum in (3.18) is attained at one of the endpoints, that is, $p = x/2$ or $p = x$. Setting $y = e^{-tx/2} \in [0, 1]$, we note that

$$\min \{y - y^2, y^2 - y^4\} = \begin{cases} y^2 - y^4, & 0 \leq y \leq \phi, \\ y - y^2, & \phi \leq y \leq 1, \end{cases}$$

where $\phi = (\sqrt{5}-1)/2$ is the golden ratio, which appears here as the root of the equation $y^2 - y^4 = y - y^2$ on $]0, 1[$. It is then easy to see that the right-hand side of (3.18), as a function of t , attains its maximum value $\phi - \phi^2 = \sqrt{5} - 2$ at $t(x) = 2x^{-1} \log(1/\phi) \rightarrow \infty$ ($x \downarrow 0$). Hence $V(t(x)) \geq (\sqrt{5} - 2)\Delta\nu(x)$, and the first inequality in (3.17) follows. \square

Our next goal is to characterize the link between the upper (lower) bounds on the values of the function $\Delta\nu(x)$ (for small x) and the lagged frequency ratios p_{j+k}/p_j (for large j) with regard to the threshold value $1/2$.

Lemma 3.9. *For a given positive integer k , the bound*

$$\Delta\nu(x) \leq k \quad (3.19)$$

is valid for all sufficiently small $x > 0$ if and only if the condition

$$\frac{p_{j+k}}{p_j} \leq \frac{1}{2} \quad (3.20)$$

is satisfied for all sufficiently large j . The similar assertion holds true when the sign \leq in both (3.19) and (3.20) is replaced by \geq .

Proof. The first part of the lemma (i.e., with \leq) is just a reformulation of definitions (see (3.13)). Indeed, applying (3.19) with $x = p_j$ implies $p_{j+k} \leq p_j/2$, which is (3.20). Conversely, if $p_j \leq x < p_{j-1}$ then by (3.20) we have $p_{j+k} \leq p_j/2 \leq x/2$, and hence $\Delta\nu(x) = \nu]x/2, x] \leq k$ as required by (3.19).

The “mirror” part (i.e., with \geq) needs a bit more care. First, note that it suffices to prove the “only if” statement in the case where $p_j > p_{j+1}$, for if $p_j = p_r$ ($r > j$) then $p_{j+k}/p_j \geq p_{r+k}/p_r$. Now, if $x \in [p_{j+1}, p_j[$ then the condition $\Delta\nu(x) \geq k$ implies that $p_{j+k} > x/2$, whence by letting $x \uparrow p_j$ we get $p_{j+k} \geq p_j/2$. Similarly, the “if” part follows by noting that $p_{j+k} \geq p_j/2$ implies $\Delta\nu(x) \geq k$ for each $x \in [p_{j+1}, p_j[$. \square

3.6 Refined asymptotic estimates

By Lemma 3.9 and the inequality (3.17), the upper bound (3.19) implies $\bar{v} \leq \bar{w} \leq k$. In some cases, however, such an estimate may not be sharp, as the next example demonstrates.

Example 3.10. Let $p_j = j2^{-j} \in \text{RT}_{1/2}$, so that by Theorem 1.1 we have $\lim_{t \rightarrow \infty} V(t) = 1$. On the other hand, (3.20) holds starting from $k = 2$, which leads to the crude bound $\bar{v} \leq 2$. An inspection shows that $\Delta\nu(\cdot) = 1$ on $[2p_{i+1}, p_{i-1}[$ and $\Delta\nu(\cdot) = 2$ on $[p_i, 2p_{i+1}[$ ($i \geq 4$). For a given $x \in [p_j, p_{j-1}[$, “excess” over the value 1 on the interval $]0, x]$ occurs on a set of total Lebesgue’s measure bounded by $\sum_{i \geq j} (2p_{i+1} - p_i) = \sum_{i \geq j} 2^{-i} = 2^{-j+1}$, which is small as compared to $x \geq p_j = j2^{-j}$ ($j \rightarrow \infty$).

This example suggests the following refinement of Lemma 3.9.

Lemma 3.11. *If for some $k \in \mathbb{N}$ the frequencies (p_j) satisfy*

$$\limsup_{j \rightarrow \infty} \frac{p_{j+k}}{p_j} \leq \frac{1}{2}, \quad (3.21)$$

then $\limsup_{t \rightarrow \infty} V(t) \leq k$. The assertion remains valid when the symbols \leq and \limsup are simultaneously replaced by \geq and \liminf .

Proof. Let us prove the first part of the lemma (with \leq and \limsup). It suffices to assume that $k = 1$, as the general case would then follow by the additivity argument (see Remark 2.4). According to (3.21) (with $k = 1$), for any $\varepsilon \in]0, 1/5]$ and all sufficiently large i we have $p_{i+1}/p_i \leq 1/2 + \varepsilon < 1$. Hence, $p_{i+1}/p_{i-1} \leq (1/2 + \varepsilon)^2 \leq 49/100 < 1/2$, and Lemma 3.9 implies that $\Delta\nu(x) \leq 2$ for all sufficiently small x .

On the other hand, the definition (3.13) of the function $\Delta\nu(\cdot)$ implies that for $u \in [p_i, p_{i-1}[$ one has $\Delta\nu(u) \leq 1$, unless $p_i < 2p_{i+1}$ ($< p_{i-1}$) and $u \in [p_i, 2p_{i+1}[$. Therefore, on each interval $[p_i, p_{i-1}[$ the value $\Delta\nu(u) = 2$ may only occur on a subset with Lebesgue’s measure not exceeding $\max\{2p_{i+1} - p_i, 0\} \leq 2\varepsilon p_i$. Hence, for a given $x \in [p_j, p_{j-1}[$ we have

$$\begin{aligned} D(x) - x &= \int_0^x (\Delta\nu(u) - 1) du \\ &= \int_{p_j}^x (\Delta\nu(u) - 1) du + \sum_{i > j} \int_{p_i}^{p_{i-1}} (\Delta\nu(u) - 1) du \\ &\leq 2\varepsilon \sum_{i \geq j} p_i \leq 2\varepsilon p_j \sum_{\ell=0}^{\infty} \left(\frac{1}{2} + \varepsilon\right)^\ell = \frac{4\varepsilon p_j}{1 - 2\varepsilon}. \end{aligned}$$

It follows that

$$\frac{D(x)}{x} - 1 \leq \frac{4\varepsilon p_j}{(1-2\varepsilon)x} \leq \frac{4\varepsilon}{1-2\varepsilon} \rightarrow 0 \quad (\varepsilon \rightarrow 0),$$

and hence $\limsup_{x \downarrow 0} D(x)/x \leq 1$. Finally, applying to (3.16) the “lim sup” part of Fatou’s lemma [16, §IV.2], we obtain

$$\limsup_{t \rightarrow \infty} \int_0^\infty e^{-y} y \frac{D(y/t)}{y/t} dy \leq \int_0^\infty e^{-y} y dy = 1,$$

and the first half of the lemma is proved.

For the second half (with \geq and \liminf), suppose again, without loss of generality, that $k = 1$. Then, according to (3.21), for any $\varepsilon \in]0, 1/2[$ and all sufficiently large i , we have $p_{i+1}/p_i \geq 1/2 - \varepsilon$. By the definition (3.13), $\Delta\nu(u) \geq 1$ for $u \in [p_{i+1}, p_i[$, unless $2p_{i+1} < p_i$ and $u \in [2p_{i+1}, p_i[$ (in which case $\Delta\nu(u) = 0$). Therefore, for a given $x \in [p_{j+1}, p_j[$ we obtain

$$\begin{aligned} x - D(x) &= \int_{p_{j+1}}^x (1 - \Delta\nu(u)) du + \sum_{i>j} \int_{p_{i+1}}^{p_i} (1 - \Delta\nu(u)) du \\ &\leq \sum_{i \geq j} (p_i - 2p_{i+1}) \mathbf{1}\{2p_{i+1} < p_i\} \\ &\leq 2\varepsilon \sum_{i \geq j} p_i \mathbf{1}\{2p_{i+1} < p_i\}. \end{aligned} \tag{3.22}$$

On account of the condition under the indicator function and using that the sequence (p_i) is nonincreasing, we note that each nonzero summand in (3.22) is at least twice as large as the next one. Hence, the sum on the right-hand side of (3.22) is dominated by

$$p_j \sum_{\ell=0}^{\infty} \left(\frac{1}{2}\right)^\ell = 2p_j. \tag{3.23}$$

Combining the estimates (3.22) and (3.23), we obtain

$$1 - \frac{D(x)}{x} \leq \frac{4\varepsilon p_j}{x} \leq \frac{4\varepsilon p_j}{p_{j+1}} \leq \frac{8\varepsilon}{1-2\varepsilon} \rightarrow 0 \quad (\varepsilon \rightarrow 0),$$

which implies that $\liminf_{x \downarrow 0} D(x)/x \geq 1$. It remains to use Fatou’s lemma in (3.16) to conclude that $\liminf_{t \rightarrow \infty} V(t) \geq 1$. \square

Corollary 3.12. *Suppose that the condition (1.2) is satisfied for some $k \in \mathbb{N}$, that is, $p_{j+k}/p_j \rightarrow 1/2$ as $j \rightarrow \infty$. Then $V(t) \rightarrow k$ as $t \rightarrow \infty$.*

Proof. Readily follows by combining the two halves of Lemma 3.11. \square

Note that Corollary 3.12 is exactly the “if” part of Theorem 1.1. In Section 5 below, where the issue of converging variance is considered in detail, we will give a direct, shorter proof of the necessity of the condition (1.2).

Example 3.13. Note that a converse statement to either half of Lemma 3.11 is *not valid*. Indeed, if $(p_j) \in \text{RT}_q$ with $q \in [0, 1/2[$, then $\Delta\nu(\cdot) = 1$ on $[p_j, 2p_j[$ and $\Delta\nu(\cdot) = 0$ on $[2p_j, p_{j-1}[$ (for j large enough). This implies that the graph $y = D(x)/x$ consists of arcs of hyperbolas with alternating monotonicity (supported on intervals of the form $[p_j, 2p_j[$ and $[2p_j, p_{j-1}[$), and in particular

$$\begin{aligned} \max_{x \in [p_j, 2p_j[} \frac{D(x)}{x} &= \frac{D(2p_j)}{2p_j} = \frac{1}{2p_j} \sum_{i \geq j} p_i = \frac{1 + \rho_j}{2}, \\ \min_{x \in [2p_j, p_{j-1}[} \frac{D(x)}{x} &= \frac{D(p_{j-1})}{p_{j-1}} = \frac{1}{p_{j-1}} \sum_{i \geq j} p_i = q(1 + \rho_j), \end{aligned} \quad (3.24)$$

where $\rho_j = p_j^{-1} \sum_{i > j} p_i$ (cf. (1.7)). The RT_q -condition implies that $\rho_j \rightarrow q/(1 - q)$ as $j \rightarrow \infty$, so from (3.24) we get

$$\frac{q}{1 - q} \leq \liminf_{x \downarrow 0} \frac{D(x)}{x} \leq \limsup_{x \downarrow 0} \frac{D(x)}{x} \leq \frac{1}{2(1 - q)}. \quad (3.25)$$

In particular, setting $q = 0$ (e.g., when (p_j) is a Poisson distribution) and taking a “doubled” sequence (i.e., determined by $\nu(dx) = \sum_{j=1}^{\infty} 2\delta_{p_j}(dx)$), by the additivity argument we get $\limsup_{t \rightarrow \infty} V(t) \leq 2 \cdot (1/2) = 1$, while $\limsup_{j \rightarrow \infty} p_{j+1}/p_j = 1$. Likewise, choosing $q = 1/3$ and again doubling the sequence, from (3.25) and by Fatou’s lemma applied to (3.16), we obtain that $\liminf_{t \rightarrow \infty} V(t) \geq 2 \cdot (1/2) = 1$, whereas $\liminf_{j \rightarrow \infty} p_{j+1}/p_j = q = 1/3$.

3.7 Proof of Theorem 1.2

We are now in a position to prove Theorem 1.2, and let us start by proving its poissonized version. By Lemma 3.8, the conditions $\bar{v} < \infty$ and $\bar{w} < \infty$ are equivalent, and according to the first half of Lemma 3.9, the latter condition holds if and only if (3.20) is satisfied for some $k \in \mathbb{N}$, which is equivalent to (1.6) (possibly, with a bigger k).

The second part of the theorem (leading to the estimate $\bar{v} \leq k$) is settled by Lemma 3.11, since condition (3.21) of the lemma coincides with condition (1.6) of the theorem.

Furthermore, by (2.16) the condition $\limsup_{n \rightarrow \infty} V_n < \infty$ is equivalent to $\bar{v} < \infty$, in which case also $\limsup_{n \rightarrow \infty} V_n = \bar{v}$ by Corollary 3.3.

Finally, the optimality of the bound $\bar{v} \leq k$ follows by merging k geometric

sequences with ratio $q = 1/2$ each and using the additivity argument (alternatively, one can consider the geometric frequencies with ratio $q = 2^{-1/k}$).

Thus, the proof of Theorem 1.2 is complete.

3.8 Comment on the threshold constant

Let us remark that the threshold $1/2$ in Theorem 1.2 is chosen to match neatly with Theorem 1.1. Replacing $1/2$ in (1.6) by some other value $0 < q < 1$ would lead to a more sophisticated upper bound

$$\limsup_{n \rightarrow \infty} V_n \leq k \lceil \log_{1/q} 2 \rceil, \quad (3.26)$$

where $\lceil x \rceil := \min \{m \in \mathbb{Z} : m \geq x\}$ is the ceiling integer part of x . Indeed, iterating the condition $\limsup_{j \rightarrow \infty} p_{j+k}/p_j \leq q$, we get $\limsup_{j \rightarrow \infty} p_{j+ik}/p_j \leq q^i \leq 1/2$, provided that $i \geq \lceil \log_{1/q} 2 \rceil$, and (3.26) follows by Lemma 3.11.

In fact, the constant $\lceil \log_{1/q} 2 \rceil$ here has the meaning of an upper bound for $\limsup_{n \rightarrow \infty} V_n$ in the geometric case with ratio q . Note that the representation (1.5) leads to a similar (in general, slightly better) estimate $\limsup_{n \rightarrow \infty} V_n \leq k(\log_{1/q} 2 + \max \delta_V(\cdot))$ (cf. (3.26)).

4 Convergence to infinity

In this section, we establish new sufficient conditions in order that $V(t) \rightarrow \infty$ as $t \rightarrow \infty$ (which, in view of (2.16), is equivalent to $V_n \rightarrow \infty$ as $n \rightarrow \infty$). Note that the combination of Theorems 4.1 and 4.3 (to be proved in Sections 4.1 and 4.2, respectively) along with discussion in Section 4.4 will settle Theorem 1.3 stated in the Introduction.

4.1 First set of sufficient conditions

It is natural to seek a condition for $V(t) \rightarrow \infty$ based on the representation (2.13), that is, in terms of the function $\Delta\nu(x)$. In turn, such a condition may be transformed into the information about the lagged ratio p_{j+k}/p_j (cf. Theorem 1.2).

Theorem 4.1. *The condition*

$$\lim_{x \downarrow 0} \Delta\nu(x) = \infty \quad (4.1)$$

implies that

$$\forall k \in \mathbb{N}, \quad \liminf_{j \rightarrow \infty} \frac{p_{j+k}}{p_j} \geq \frac{1}{2}, \quad (4.2)$$

which in turn implies that $V(t) \rightarrow \infty$ as $t \rightarrow \infty$.

Proof. If condition (4.1) holds then for any $k \in \mathbb{N}$ we have $\Delta\nu(x) \geq k$ for all sufficiently small $x > 0$. By Lemma 3.9, this implies that $p_{j+k}/p_j \geq 1/2$ for all j large enough, and (4.2) follows. Further, condition (4.2) implies convergence of $V(t)$ to infinity by Lemma 3.11. \square

Note that condition (4.2) is obviously fulfilled for any sequence (p_j) from RT_1 , in which case it is well known that $V(t) \rightarrow \infty$ [13,25]. The next example demonstrates that there are instances of frequencies (p_j) satisfying (4.1) but *not* in RT_1 . This example will also show that conditions (4.1) and (4.2) of Theorem 4.1 are *not necessary* in order that $V(t) \rightarrow \infty$.

Example 4.2. Let $0 < q < 1$ and suppose that the sequence (p_j) consists of the values q^i , each repeated i times ($i = 1, 2, \dots$), which corresponds to the measure $\nu(dx) = \sum_{i=1}^{\infty} i \delta_{q^i}(dx)$. Note that the sequence (p_j) is not in any RT -class, since $\limsup_{j \rightarrow \infty} p_{j+1}/p_j = 1$ but $\liminf_{j \rightarrow \infty} p_{j+1}/p_j = q$. However, for any $q \in]0, 1[$ we have $V(t) \rightarrow \infty$, since for $t \in [q^{-j}, q^{-j-1}]$

$$\begin{aligned} V(t) &= \sum_{i=1}^{\infty} i (e^{-q^i t} - e^{-2q^i t}) \geq j (e^{-q^j t} - e^{-2q^j t}) \\ &\geq j \min_{y \in [1, q^{-1}]} (e^{-y} - e^{-2y}) = j (e^{-1/q} - e^{-2/q}) \rightarrow \infty \quad (j \rightarrow \infty). \end{aligned}$$

If $1/2 \leq q < 1$ then for $x \in [q^j, q^{j-1}[$ we have $\Delta\nu(x) \geq j \rightarrow \infty$ as $x \downarrow 0$, and condition (4.1) is valid. On the other hand, if $0 < q < 1/2$ then $\Delta\nu(x) = 0$ for $x \in [2q^j, q^{j-1}[$, hence $\liminf_{x \downarrow 0} \Delta\nu(x) = 0$ and (4.1) fails. Also, for any $k \geq 1$, we have $\liminf_{j \rightarrow \infty} p_{j+k}/p_j = q < 1/2$, so condition (4.2) is not valid.

4.2 Another set of conditions

A different sufficient condition exploits the link between $V(t)$ and the mean number of singleton boxes $\Phi_1(t)$, as in Lemma 3.4. An equivalent condition may be set in terms of the tail ratio $\rho_j = p_j^{-1} \sum_{i>j}^{\infty} p_i$ (see (1.7)). Recall the definition (3.7) of the measure $\tilde{\nu}$.

Theorem 4.3. *The condition*

$$\lim_{x \downarrow 0} \frac{\tilde{\nu}[0, x]}{x} = \infty \quad (4.3)$$

is equivalent to

$$\lim_{j \rightarrow \infty} \rho_j = \infty, \quad (4.4)$$

and each one implies that $V(t) \rightarrow \infty$ as $t \rightarrow \infty$.

Proof. By the estimate (3.12), condition (4.3) implies $\Phi_1(t) \rightarrow \infty$, which is equivalent to $V(t) \rightarrow \infty$ by (3.1). So it remains to show that (4.3) and (4.4) are equivalent to each other. Observe that for $p_{j+1} \leq x < p_j$ we have $x^{-1} \tilde{\nu}[0, x] \geq \rho_j$, hence (4.4) implies (4.3). To prove the converse, note that if $p_{j+1} = p_j$ then $\rho_j = 1 + \rho_{j+1}$, so it suffices to consider the case where $p_{j+1} < p_j$. Then

$$\rho_j = \inf_{p_{j+1} \leq x < p_j} \frac{\tilde{\nu}[0, x]}{x} \rightarrow \infty \quad (j \rightarrow \infty),$$

when the condition (4.3) holds, and hence (4.4) follows. \square

4.3 A counterexample to Theorem 4.3

We construct here an example demonstrating that conditions (4.3), (4.4) are *not necessary* in order that $V(t) \rightarrow \infty$ (or, equivalently, $\Phi_1(t) \rightarrow \infty$). In particular, due to the estimate (3.10) (with $r = 2$), this example will show that $V(t) \rightarrow \infty$ does not necessarily imply $\Phi_2(t) \rightarrow \infty$. On the other hand, in view of the inequality

$$2\Phi_2(t) \geq \sum_{1/2 < tp_j \leq 1} (tp_j)^2 e^{-tp_j} \geq \frac{e^{-1}}{4} \#\{p_j \in]1/(2t), 1/t]\} = \frac{e^{-1}}{4} \Delta\nu(1/t),$$

it is a priori clear that $\Phi_2(t)$ cannot be uniformly bounded in such a situation, because $\bar{w} = \infty$ according to (3.17).

Example 4.4. Let k_0, k_1, k_2, \dots be an increasing integer sequence. Take the frequencies (p_j) in the form

$$p_j = \begin{cases} k_1^{-1}, & 0 < j \leq k_0, \\ k_{i+1}^{-1}, & k_0 + \dots + k_{i-1} < j \leq k_0 + \dots + k_i, \end{cases} \quad (4.5)$$

which corresponds to the measure

$$\nu(dx) = \sum_{j=1}^{\infty} \delta_{p_j}(dx) = \sum_{i=0}^{\infty} k_i \delta_{k_{i+1}^{-1}}(dx). \quad (4.6)$$

That is to say, the array of boxes is partitioned in blocks so that i -th block contains k_i boxes of frequencies $1/k_{i+1}$ ($i = 0, 1, 2, \dots$).

The heuristics underlying this example is as follows. A prototype instance is a block of k equal boxes each with frequency, say, q . The mean number of singleton boxes within the block is a single-wave function $ktq e^{-tq}$ which increases to its maximum k/e at time $t = 1/q$ and then goes down to 0. Now, the idea is to combine a series of such blocks in order to guarantee a suitable overlap of the waves produced by successive blocks. If the sequence (k_i) grows fast enough, then for each $i = 0, 1, 2, \dots$ there exists a time instant (of order of k_{i+1}) when boxes belonging to i -th block start to get occupied. After some time, the mean number of singletons among these boxes is still relatively large, say not less than $\log \log k_i$, but the expected number of balls that fall in boxes of further blocks becomes large too, and almost all these balls produce singleton boxes, since k_{i+1} is yet much larger (hence the frequencies are smaller). As time passes, all boxes belonging to blocks $0, 1, \dots, i$ are likely to contain more than one ball each, while the balls hitting other blocks remain sole representatives of their boxes.

To make this heuristic work, we choose

$$k_i := 2^{2^i}, \quad i = 0, 1, 2, \dots, \quad (4.7)$$

so that $k_{i+1} = k_i^2$ for all i . We wish to check that $\Phi_1(t)$ goes to infinity but $\Phi_2(t)$ does not. Using (2.10) and (4.6) we have

$$\Phi_1(t) = t \int_0^\infty x e^{-tx} \nu(dx) = \sum_{i=0}^\infty \frac{tk_i}{k_{i+1}} e^{-t/k_{i+1}} =: \sum_{i=0}^\infty A_i(t), \quad (4.8)$$

$$\Phi_2(t) = \frac{t^2}{2} \int_0^\infty x^2 e^{-tx} \nu(dx) = \frac{1}{2} \sum_{i=0}^\infty \frac{t^2 k_i}{k_{i+1}^2} e^{-t/k_{i+1}} =: \frac{1}{2} \sum_{i=0}^\infty B_i(t). \quad (4.9)$$

As a function of t , each summand $A_i(t)$ in the sum (4.8) increases up to the maximum value $A_i(t_i^*) = k_i e^{-1}$ attained at $t_i^* = k_{i+1}$, and then decreases to zero. Two consecutive summands, $A_i(t)$ and $A_{i+1}(t)$, are equal at the point

$$t'_i := \frac{k_{i+1}^2}{k_{i+1} - 1} \log k_i,$$

where their common value is

$$A_i(t'_i) = \frac{k_{i+1}}{k_{i+1} - 1} k_i^{-1/(k_{i+1}-1)} \log k_i.$$

Using the elementary inequality $k^{-1/(k-1)} \geq e^{-1}$ ($k > 1$), we note that

$$A_i(t'_i) \geq k_i^{-1/(k_i-1)} \log k_i \geq e^{-1} \log k_i.$$

Since $t'_{i-1} < t_i^* < t'_i$ ($i = 1, 2, \dots$), it follows that for all $t \in [t'_{i-1}, t'_i]$,

$$\Phi_1(t) \geq A_i(t) \geq e^{-1} \log k_{i-1},$$

hence

$$\liminf_{t \rightarrow \infty} \Phi_1(t) \geq e^{-1} \liminf_{i \rightarrow \infty} \log k_{i-1} = \infty.$$

Turning to $\Phi_2(t)$, note that the summand $B_i(t)$ in (4.9) attains its maximum value at the point $t = 2t_i^* = 2k_{i+1}$ and $B_i(2t_i^*) = 4 e^{-2} k_i$, so

$$\Phi_2(2t_i^*) \geq B_i(2t_i^*) = 4 e^{-2} k_i \rightarrow \infty \quad (i \rightarrow \infty).$$

On the other hand, on the sequence $t_j'' := 3k_{j+1} \log k_j$ one has

$$B_i(t_j'') = \frac{(t_j'')^2 k_i}{k_{i+1}^2} e^{-t_j''/k_{i+1}} = \frac{9k_{j+1}^2 \log^2 k_j}{k_{i+1}^{3/2}} \exp\left(-\frac{3k_{j+1} \log k_j}{k_{i+1}}\right).$$

Setting $x = k_{i+1}$ and $a = k_{j+1} \log k_j$, we note that the function $x^{-3/2} e^{-3a/x}$ increases for $0 < x \leq 2a$. Hence, for all $i = 0, 1, \dots, j$,

$$B_i(t_j'') \leq B_j(t_j'') = \frac{9 \log^2 k_j}{k_j^2},$$

and therefore

$$\sum_{i=0}^j B_i(t_j'') \leq (j+1) B_j(t_j'') = \frac{9(j+1) \log^2 k_j}{k_j^2}. \quad (4.10)$$

For $i \geq j+1$, we have

$$B_i(t_j'') \leq \frac{9k_{j+1}^2 \log^2 k_j}{k_i k_{i+1}} \leq \frac{9 \log^2 k_j}{k_i},$$

and since $k_i = 2^{2^i} \geq 2^{4i}$ for $i \geq 4$, it follows

$$\sum_{i=j+1}^{\infty} B_i(t_j'') \leq 9 \cdot 2^{2j} \sum_{i=j+1}^{\infty} 2^{-4i} = \frac{3}{5 \cdot 2^{2j}}. \quad (4.11)$$

Combining the estimates (4.10) and (4.11) yields $\Phi_2(t_j'') \rightarrow 0$ as $j \rightarrow \infty$.

Thus $\Phi_2(t)$ does not have a limit as $t \rightarrow \infty$, and moreover

$$\liminf_{t \rightarrow \infty} \Phi_2(t) = 0, \quad \limsup_{t \rightarrow \infty} \Phi_2(t) = \infty.$$

Finally, it is easy to see directly that in this example the limit in (4.4) does not exist. Indeed, along the subsequence $j = k_0 + k_1 + \dots + k_i$, according to (4.5) and (4.7),

$$\rho_j = k_{i+1} \left(\frac{k_{i+1}}{k_{i+2}} + \frac{k_{i+2}}{k_{i+3}} + \dots \right) = 1 + O(k_{i+1}^{-1}) \rightarrow 1 \quad (i \rightarrow \infty).$$

On the other hand, for $j = k_0 + k_1 + \cdots + k_i + 1$ we have

$$\rho_j = k_{i+2} \left(\frac{k_{i+1} - 1}{k_{i+2}} + \frac{k_{i+2}}{k_{i+3}} + \cdots \right) \geq k_{i+1} - 1 \rightarrow \infty \quad (i \rightarrow \infty).$$

Karlin [25, page 384] gives an example of frequencies for which $V(t)$ converges to 0 along a sequence of values of t , and converges to ∞ along another sequence; in that case $\Phi_1(t)$ demonstrates the same type of behavior. Our Example 4.4 exhibits a more exotic “second order” pathology: this time, $\Phi_1(t) \rightarrow \infty$ but $\Phi_2(t)$ oscillates between 0 and ∞ .

4.4 Relationship between the various sufficient conditions

First of all, note that condition (4.2) in Theorem 4.1 does not imply condition (4.1). A counterexample may be constructed by a slight modification of Example 4.2 as follows: define the frequencies (p_j) by setting $\nu(dx) = \sum_{i=1}^{\infty} i \delta_{\tilde{p}_i}$, where $\tilde{p}_i := i^{-1}2^{-i}$, then $\liminf_{j \rightarrow \infty} p_{j+k}/p_j = \liminf_{i \rightarrow \infty} \tilde{p}_{i+1}/\tilde{p}_i = 1/2$ (so that (4.2) is satisfied), but for $(i+1)^{-1}2^{-i} \leq x < i^{-1}2^{-i}$ we have $\Delta\nu(x) = 0$, hence $\liminf_{x \downarrow 0} \Delta\nu(x) = 0$ and (4.1) fails.

Further, it is easy to see that condition (4.2) in Theorem 4.1 implies the set of equivalent conditions (4.3), (4.4) in Theorem 4.3, but not the other way around. Indeed, if (4.2) is satisfied then for ρ_j defined in (1.7) we have

$$\liminf_{j \rightarrow \infty} \rho_j \geq \liminf_{j \rightarrow \infty} \sum_{k=1}^M \frac{p_{j+k}}{p_j} \geq M \cdot \frac{1}{2} \rightarrow \infty \quad (M \rightarrow \infty),$$

and condition (4.4) follows. On the other hand, we have seen that in Example 4.2 condition (4.2) fails, while for $q^j \leq x < q^{j-1}$ we have

$$\frac{\tilde{\nu}[0, x]}{x} = \frac{1}{x} \sum_{i \geq j} i q^i \geq \frac{j}{q^{j-1}} \sum_{i \geq j} q^i = \frac{j q}{1 - q} \rightarrow \infty \quad (j \rightarrow \infty),$$

and the condition (4.3) is valid.

As Example 4.4 shows, a converse to Theorem 4.3 is *not valid*, unless under further assumptions on the measure $\tilde{\nu}$ (cf. [13,25]). For instance, if $\tilde{\nu}[0, x]$ varies regularly at zero, then Karamata’s Tauberian theorem (see [8, §1.7.2] or [16, §XIII.5]) applied to (3.8) yields $\tilde{\nu}[0, x]/x \sim c \Phi_1(1/x)$ as $x \downarrow 0$, so that the convergence $\Phi_1(t) \rightarrow \infty$ as $t \rightarrow \infty$ does imply the condition (4.3).

Remark 4.5. By Karamata’s Tauberian theorem, the convergence

$$\Phi_1(t) = t \int_0^{\infty} e^{-tx} \tilde{\nu}(dx) \rightarrow c \quad (t \rightarrow \infty)$$

is equivalent to $\tilde{\nu}[0, x]/x \rightarrow c$ as $x \downarrow 0$. Interestingly, the implication may fail for $c = \infty$, as Example 4.4 demonstrates.

5 Convergence to a finite limit

We will now investigate the situation where the variance $V(t)$ has a finite limit as $t \rightarrow \infty$, which is the central topic of this work (see Theorem 1.1). As already mentioned in Section 3.6, the “if” part of Theorem 1.1 follows from Corollary 3.12. So the main goal of this section is to prove the “only if” part (i.e., the sufficiency of the condition (1.2)), but we will also give a streamlined proof of the necessity.

5.1 Criterion of convergence

Recall that $D(\cdot)$ is a primitive function of $\Delta\nu(\cdot)$, defined by (3.15).

Lemma 5.1. *In order that there exist a finite limit*

$$\lim_{t \rightarrow \infty} V(t) =: v, \quad (5.1)$$

it is necessary and sufficient that

$$\lim_{x \downarrow 0} \frac{D(x)}{x} = v. \quad (5.2)$$

Proof. Note that, according to (3.6), $v > 0$. By the representation (3.14), we can rewrite (5.1) as

$$\int_0^\infty e^{-tx} \, dD(x) \sim \frac{v}{t} \quad (t \rightarrow \infty). \quad (5.3)$$

By Karamata’s Tauberian theorem (see [8, §1.7.2], [16, §XIII.5]), the relation (5.3) is equivalent to $D(x) \sim vx$ as $x \downarrow 0$, which is the same as (5.2). \square

5.2 Some implications of convergence

Lemma 5.2. *Suppose that the limit (5.2) exists, and let $\alpha, \beta > 0$ be arbitrary variables such that $\alpha, \beta \downarrow 0$ and $(\alpha + \beta)/(\beta - \alpha) = O(1)$. Then*

$$\lim_{\alpha, \beta \downarrow 0} \frac{D(\beta) - D(\alpha)}{\beta - \alpha} = v.$$

Proof. Using (5.2), we have

$$\frac{D(\beta) - D(\alpha)}{\beta - \alpha} = \frac{v\beta(1 + o(1)) - v\alpha(1 + o(1))}{\beta - \alpha} = v + \frac{o(1)(\alpha + \beta)}{\beta - \alpha} \rightarrow v,$$

since the ratio $(\alpha + \beta)/(\beta - \alpha)$ is bounded. \square

Lemma 5.3. *If the finite limit (5.1) exists then the limiting value v must be a positive integer number, $v = k \in \mathbb{N}$, and in this case*

$$\lim_{x \downarrow 0} \frac{\lambda\{u \in]0, x] : \Delta\nu(u) \neq k\}}{x} = 0, \quad (5.4)$$

where $\lambda\{\cdot\}$ denotes Lebesgue's measure on \mathbb{R}_+ .

Proof. By Lemma 3.8, the function $\Delta\nu(u)$ is uniformly bounded. By definition, it counts the number of frequencies p_j in the interval $]u/2, u]$, therefore $\Delta\nu(u)$ is piecewise constant, with jumps at points $u = p_j$ and $u = 2p_j$. Thus, for any given interval $]x/2, x]$ the total number of such jumps is uniformly bounded by a constant, say $M < \infty$.

Let $]\alpha, \beta[$ be the maximal open subinterval of $]x/2, x]$, on which $\Delta\nu(\cdot)$ is constant. Clearly, its length satisfies $\beta - \alpha \geq x/2(M + 1)$, thus

$$0 \leq \frac{\alpha + \beta}{\beta - \alpha} \leq \frac{2x}{x/2(M + 1)} = 4(M + 1). \quad (5.5)$$

Consider a closed interval $[\alpha_1, \beta_1] \subset]\alpha, \beta[$ with $\alpha_1 = (3\alpha + \beta)/4$, $\beta_1 = (3\beta + \alpha)/4$. Since $\alpha_1 + \beta_1 = \alpha + \beta$ and $\beta_1 - \alpha_1 = (\beta - \alpha)/2$, by the bound (5.5) Lemma 5.2 applies to yield

$$\frac{1}{\beta_1 - \alpha_1} \int_{\alpha_1}^{\beta_1} \Delta\nu(u) \, du = \frac{D(\beta_1) - D(\alpha_1)}{\beta_1 - \alpha_1} \rightarrow v \quad (x \downarrow 0). \quad (5.6)$$

But the function $\Delta\nu(\cdot)$ is constant on $]\alpha, \beta[\supset]\alpha_1, \beta_1]$, hence its sole (integer) value must coincide with the asymptotic mean v given by (5.6). In particular, v must be integer, $v = k \in \mathbb{N}$.

Along the same lines, one can show that for any $\varepsilon > 0$ and all small enough x , the function $\Delta\nu(\cdot)$ takes the value $v = k$ on the interval $]x/2, x]$ everywhere except on a set of Lebesgue's measure smaller than εx . Thus, Lebesgue's measure of the set $\{u \in]0, x] : \Delta\nu(u) \neq k\}$ is bounded by $\varepsilon \sum_{i=1}^{\infty} 2^{-i+1}x = 2\varepsilon x$, and since ε is arbitrary, (5.4) follows. \square

5.3 Lagged frequency ratio and the proof of Theorem 1.1

Lemma 5.4. *If the limit (5.1) exists (hence $v = k \in \mathbb{N}$ by Lemma 5.3), then (cf. (1.2))*

$$\lim_{j \rightarrow \infty} \frac{p_{j+k}}{p_j} = \frac{1}{2}. \quad (5.7)$$

Proof. Without loss of generality, it suffices to consider $j \in \mathbb{N}$ such that $2p_{j+k} \neq p_j$. Suppose first that $2p_{j+k} < p_j$. Then for $x \in [2p_{j+k}, p_j[$ we have $]x/2, x] \subset]p_{j+k}, p_j[$ and hence $\Delta\nu(x) \leq k - 1$. Therefore,

$$D(p_j) - D(2p_{j+k}) = \int_{2p_{j+k}}^{p_j} \Delta\nu(u) du \leq (k - 1)(p_j - 2p_{j+k}). \quad (5.8)$$

Using that $D(x) = kx(1 + o(1))$ as $x \downarrow 0$ (see Lemma 5.1), from (5.8) we deduce that $\liminf_{j \rightarrow \infty} p_{j+k}/p_j \geq 1/2$, which, together with the hypothesis $p_{j+k}/p_j < 1/2$ (see above), implies (5.7).

Likewise, if $p_j < 2p_{j+k}$ then for $x \in [p_j, 2p_{j+k}[$ we have $]x/2, x] \supset [p_{j+k}, p_j]$, hence $\Delta\nu(x) \geq k + 1$ and (cf. (5.8))

$$D(2p_{j+k}) - D(p_j) = \int_{p_j}^{2p_{j+k}} \Delta\nu(u) du \geq (k + 1)(2p_{j+k} - p_j).$$

Similarly as before, this simplifies to $\limsup_{j \rightarrow \infty} p_{j+k}/p_j \leq 1/2$, and since we assumed that $p_{j+k}/p_j < 1/2$, (5.7) follows. The proof is complete. \square

Let us now show the converse of Lemma 5.4 (as mentioned at the beginning of Section 5, this also follows from Corollary 3.12).

Lemma 5.5. *Assume that the sequence (p_j) satisfies the condition (5.7) for some $k \in \mathbb{N}$. Then the limit (5.1) exists and $v = k$.*

Proof. By additivity, it suffices to prove that for each subsequence $p_j^{(i)} := p_{i+k(j-1)}$ ($i = 1, \dots, k$), its contribution to the limit (5.1) equals exactly 1. Thus the proof is reduced to showing that if $(p_j) \in \text{RT}_{1/2}$ then

$$V(t) = \sum_{j=1}^{\infty} (e^{-tp_j} - e^{-2tp_j}) \rightarrow 1 \quad (t \rightarrow \infty). \quad (5.9)$$

By the RT-condition, $2p_{j+1} = p_j(1 + \gamma_j)$, where $\gamma_j \rightarrow 0$ as $j \rightarrow \infty$. Hence, for any $\varepsilon \in]0, 1/3]$ and all j large enough we have $|\gamma_j| \leq \varepsilon$. In particular, $p_{j+2}/p_j \leq (1+\varepsilon)^2/4 \leq 4/9 < 1/2$, which implies by Lemma 3.9 that $\Delta\nu(x) \leq 2$

for small x . By Lemma 3.8 and the estimate (3.1), it follows that $\Phi_1(\cdot)$ is bounded. Returning to (5.9), observe that

$$\sum_{j=j_0}^M (e^{-tp_j} - e^{-2tp_j}) = \sum_{j=j_0}^M e^{-tp_j} (1 - e^{-tp_j \gamma_j}) - e^{-2tp_{j_0}} + e^{-2tp_{M+1}}. \quad (5.10)$$

By the inequality $|1 - e^{-y}| \leq |y| e^{|y|}$, the sum in (5.10) is dominated by

$$\sum_{j=j_0}^M e^{-tp_j(1-\varepsilon)} tp_j \varepsilon \leq \varepsilon \sum_{j=1}^{\infty} e^{-tp_j(1-\varepsilon)} tp_j = \frac{\varepsilon}{1-\varepsilon} \Phi_1(t(1-\varepsilon)) = O(\varepsilon).$$

Passing to the limit in (5.10) as $M \rightarrow \infty$, we obtain $V(t) = 1 + o(1) + O(\varepsilon)$ as $t \rightarrow \infty$, and since ε is arbitrarily small, we arrive at (5.9). \square

We are now able to complete the proof of our main Theorem 1.1 characterizing the case of converging variance. Indeed, putting together Lemmas 5.4 and 5.5 yields the desired criterion for $V(t) \rightarrow v$. Appealing to Corollary 3.3 we conclude that the same condition applies to $V_n \rightarrow v$.

5.4 Link with Karlin's condition

In conclusion, let us recall that Karlin's sufficient condition for $V(t) \rightarrow v$ [25, Theorem 2] involves (i) the condition $\limsup_{j \rightarrow \infty} p_{j+1}/p_j < 1$ and (ii) an integral condition, which in our notation reads

$$\lim_{x \rightarrow \infty} \frac{1}{x} \int_0^x \Delta\nu(1/y) dy = v, \quad (5.11)$$

or, after an obvious change of variables,

$$\lim_{x \downarrow 0} x \int_x^\infty \Delta\nu(u) u^{-2} du = v. \quad (5.12)$$

Throughout his paper, Karlin also postulates that the function $\nu_c(x) = \nu]x, \infty[$ is regularly varying at zero (see [25, pages 376–377]). As we shall see, this condition is superfluous and may be omitted (in fact, Karlin's proof of his Theorem 2 only requires the boundedness of $\Delta\nu(x)$, which follows easily from condition (i)). Note that condition (i) itself is not necessary for the convergence of $V(t)$: for instance, it does not hold for a sequence (p_j) obtained by merging several geometric sequences with ratio 1/2 into one.

Furthermore, application of condition (5.11) to the geometric case (with ratio q) yields the following (cf. [25, Example 6] containing an error). Let $\log_{1/q} 2 =$

$k + \delta$, where $k = [\log_{1/q} 2]$ is the integer part of $\log_{1/q} 2$ and $\delta \in [0, 1[$ is its fractional part. From the definition of $\Delta\nu(\cdot)$ it follows that

$$\begin{aligned} \frac{1}{x} \int_0^x \Delta\nu(1/y) dy &= \frac{1}{x} \int_0^x \left([\log_{1/q}(2y)] - [\log_{1/q} y] \right) dy \\ &= \frac{1}{x} \int_0^x \left([k + \delta + \log_{1/q} y] - [\log_{1/q} y] \right) dy \\ &= k + \frac{1}{x} \int_0^x \left([\delta + \log_{1/q} y] - [\log_{1/q} y] \right) dy. \end{aligned} \quad (5.13)$$

If $\delta = 0$, the integral in (5.13) vanishes and condition (5.11) yields $v = k$. However, if $0 < \delta < 1$ then (5.13) does not have a limit as $x \rightarrow \infty$, since for $x = q^{-j}$ the integral term amounts to

$$q^j \sum_{i=1}^j q^{-i} (1 - q^\delta) \rightarrow \frac{1 - q^\delta}{1 - q} \quad (j \rightarrow \infty),$$

whereas for $x = q^{-j-1+\delta}$ it reads

$$q^{j+1-\delta} \sum_{i=1}^j q^{-i} (1 - q^\delta) \rightarrow q^{1-\delta} \frac{1 - q^\delta}{1 - q} \quad (j \rightarrow \infty).$$

As a result, condition (5.11) is satisfied if and only if $\log_{1/q} 2 = k \in \mathbb{N}$, or equivalently $q = 2^{-1/k}$. Our Theorem 1.1 gives the same result, so (5.11) proves to yield a correct answer in the whole range of the geometric case.

This observation brings up the question about the exact relationship between Karlin's condition (5.11) (or (5.12)) and our criterion (5.2). Surprisingly enough, we can demonstrate the following.

Theorem 5.6. *Condition (5.12) is equivalent to (5.2), and hence the former is necessary and sufficient in order that $V(t) \rightarrow v$ as $t \rightarrow \infty$.*

Proof. Suppose condition (5.2) holds. Using the notation $D(x)$ (see (3.15)) and integrating by parts, we get

$$\begin{aligned} x \int_x^\infty \Delta\nu(u) \frac{du}{u^2} &= x \int_x^\infty u^{-2} dD(u) = -\frac{D(x)}{x} + 2x \int_x^\infty D(u) u^{-3} du \\ &= -\frac{D(x)}{x} + 2 \int_1^\infty \frac{D(xs)}{xs} s^{-2} ds \rightarrow -v + 2v \int_1^\infty s^{-2} ds = v \quad (x \downarrow 0), \end{aligned}$$

where we used that the function $D(u)/u$ is bounded on $]0, \infty[$ (in particular, the dominated convergence theorem can be applied). Hence, (5.12) follows.

On the other hand, condition (5.12) amounts to

$$\lim_{x \downarrow 0} xG(x) = v, \quad G(x) := \int_x^\infty \Delta\nu(u) u^{-2} du. \quad (5.14)$$

Again integrating by parts, we obtain

$$\begin{aligned} \frac{1}{x} \int_0^x \Delta\nu(u) \, du &= -\frac{1}{x} \int_0^x u^2 \, dG(u) = -xG(x) + \frac{2}{x} \int_0^x uG(u) \, du \\ &= -xG(x) + 2 \int_0^1 xsG(xs) \, ds \rightarrow -v + 2v = v \quad (x \downarrow 0), \end{aligned}$$

where we may use dominated convergence because the function $uG(u)$ is bounded on $]0, 1]$ due to (5.14). Thus, condition (5.12) implies (5.2), and the proof is complete. \square

Remark 5.7. The statement of Theorem 5.6 is a particular case of a general Karamata theorem (see [8, §1.6.3], [16, §VIII.9]), according to which the limiting relation (5.2) is equivalent to either of the limits

$$\begin{aligned} \lim_{x \downarrow 0} x^{\sigma-1} \int_x^\infty \Delta\nu(u) u^{-\sigma} \, du &= \frac{v}{\sigma-1} \quad (\sigma > 1), \\ \lim_{x \downarrow 0} x^{\sigma-1} \int_0^x \Delta\nu(u) u^{-\sigma} \, du &= \frac{v}{1-\sigma} \quad (\sigma < 1). \end{aligned}$$

(Note that (5.2) itself is contained in the second formula with $\sigma = 0$.) That is to say, our condition (5.2) may be included in a parametric family of mutually equivalent criteria, set in terms of rescaled integrals of the function $\Delta\nu(\cdot)$ against polynomial weights (the canonical criterion (5.2) being apparently the simplest). We have given a direct proof of Theorem 5.6 because of the historic interest of Karlin's condition (5.11).

Acknowledgments

Main part of this research was carried out during L. V. Bogachev's visit to the University of Utrecht in June 2006, made possible due to a grant under the Global Exchange Programme (GEP) of the Worldwide Universities Network (WUN), awarded at the University of Leeds. Hospitality of the hosts at Utrecht is much appreciated. The authors are grateful to the anonymous referee for valuable comments.

References

- [1] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*, Springer-Verlag, Berlin–Heidelberg–New York, 1989.
- [2] M. Archibald, A. Knopfmacher, H. Prodinger, The number of distinct values in a geometrically distributed sample, *European J. Combin.* 27 (2006) 1059–1081.

- [3] R. R. Bahadur, On the number of distinct values in a large sample from an infinite discrete distribution, *Proc. Nat. Inst. Sci. India Part A, Suppl. II* 26 (1960) 67–75.
- [4] R. E. Barlow, F. Proschan, *Mathematical Theory of Reliability* (with contributions by L. C. Hunter), John Wiley & Sons, New York, 1965; SIAM, Philadelphia, PA, 1996.
- [5] Yu. Baryshnikov, B. Eisenberg, G. Stengle, A necessary and sufficient condition for the existence of the limiting probability of a tie for first place, *Statist. Probab. Lett.* 23 (1995) 203–209.
- [6] J. P. Bell, S. N. Burris, Asymptotics for logical limit laws: when the growth of the components is in an RT class, *Trans. Amer. Math. Soc.* 355 (2003) 3777–3794.
- [7] C. Berg, J. P. R. Christensen, P. Ressel, *Harmonic Analysis on Semigroups*, Springer-Verlag, Berlin–Heidelberg–New York, 1984.
- [8] N. H. Bingham, C. M. Goldie, J. L. Teugels, *Regular Variation*, Cambridge Univ. Press, Cambridge, 1987.
- [9] S. N. Burris, *Number Theoretic Density and Logical Limit Laws*, Amer. Math. Soc., Providence, RI, 2001.
- [10] C. A. Charalambides, *Combinatorial Methods in Discrete Distributions*, Wiley-Interscience, Hoboken, NJ, 2005.
- [11] D. A. Darling, Some limit theorems associated with multinomial trials, in: L. M. Le Cam, J. Neyman (Eds.), *Proc. Fifth Berkeley Sympos. Math. Statist. Probab.* (Berkeley, CA, 1965/66), vol. II: Contributions to Probability Theory, Part 1, Univ. California Press, Berkeley, CA, 1967, pp. 345–350.
- [12] M. Dutko, *Limit theorems for infinite urn models in probability theory*, Ph.D. thesis, Annex (UP), Microfilm Cd5163, Pennsylvania State Univ., Philadelphia, PA, 1984.
- [13] M. Dutko, Central limit theorems for infinite urn models, *Ann. Probab.* 17 (1989) 1255–1263.
- [14] B. Eisenberg, G. Stengle, G. Strang, The asymptotic probability of a tie for first place, *Ann. Appl. Probab.* 3 (1993) 731–745.
- [15] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. I, 3rd edn., John Wiley & Sons, New York, 1968.
- [16] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. II, 2nd edn., John Wiley & Sons, New York, 1971.
- [17] A. Gnedin, J. Pitman, Moments of convex distribution functions and completely alternating sequences, in: *Festschrift for Avner Friedman (IMS Lecture Notes Monogr. Ser.)*, Inst. Math. Statist., Beachwood, OH, 2007 (to appear). Available from: [arXiv:math.PR/0602091](https://arxiv.org/abs/math.PR/0602091).

- [18] B. L. Granovsky, Asymptotics of counts of small components in random combinatorial structures and models of coagulation-fragmentation, Preprint, 2005. Available from: [arXiv:math.PR/0511381](https://arxiv.org/abs/math.PR/0511381).
- [19] P. Hitczenko, G. Louchard, Distinctness of compositions of an integer: a probabilistic analysis, *Random Structures Algorithms* 19 (2001) 407–437.
- [20] H.-K. Hwang, S. Janson, Local limit theorems for finite and infinite urn models, U.U.D.M. Report 2006:9, Uppsala Univ., Sweden, 2006. Available from: [arXiv:math.PR/0604397](https://arxiv.org/abs/math.PR/0604397).
- [21] V. A. Ivanov, G. I. Ivchenko, Yu. I. Medvedev, Discrete problems in probability theory, (Russian) in: *Itogi Nauki Tekhn. (Teor. Veroyatnost. Mat. Statist. Teor. Kibernet., vol. 22)*, Vsesoyuz. Inst. Nauchn. i Tekhn. Inform. (VINITI), Moscow, 1984, pp. 3–60; (English translation) *J. Soviet Math.* 31 (1985) 2759–2795.
- [22] P. Jacquet, W. Szpankowski, Analytical dePoissonization and its applications, *Theoret. Comput. Sci.* 201 (1998) 1–62.
- [23] S. Janson, Rounding of continuous random variables and oscillatory asymptotics, *Ann. Probab.* 34 (2006) 1807–1826.
- [24] N. L. Johnson, S. Kotz, *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*, John Wiley & Sons, New York, 1977.
- [25] S. Karlin, Central limit theorems for certain infinite urn schemes, *J. Math. Mech.* 17 (1967) 373–401.
- [26] S. V. Kerov, Coherent allocations, and the Ewens-Pitman formula, (Russian) Preprint, POMI 21/1995, 1995, pp. 1–15; (English translation) Coherent random allocations, and the Ewens-Pitman formula (edited and with comments by A. Gnedin), in: A. M. Vershik (Ed.), *Representation Theory, Dynamical Systems, Combinatorial and Algorithmic Methods; Part 12*, *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, vol. 325, 2005, pp. 127–145.
- [27] V. F. Kolchin, B. A. Sevast'yanov, V. P. Chistyakov, *Random Allocations*, (Russian) Izdat. "Nauka", Moscow, 1976; (English translation) V. H. Winston & Sons, Washington, D.C.; Halsted Press (John Wiley & Sons), New York, 1978.
- [28] S. Kotz, N. Balakrishnan, Advances in urn models during the past two decades, in: N. Balakrishnan (Ed.), *Advances in Combinatorial Methods and Applications to Probability and Statistics*, Birkhäuser, Boston, 1997, pp. 203–257.
- [29] G. Louchard, H. Prodinger, M. D. Ward, The number of distinct values of some multiplicity in sequences of geometrically distributed random variables, in: C. Martínez (Ed.), *2005 International Conference on Analysis of Algorithms, Discrete Math. Theor. Comput. Sci. (DMTCS) Proc., AD (electronic)*, Assoc. DMTCS, Nancy, 2005, pp. 231–256. Available from: <http://www.dmtcs.org/pdfpapers/dmAD0122.pdf>.

- [30] V. B. Nevzorov, *Records: Mathematical Theory*, (Russian) Izdat. FAZIS, Moscow, 2000; (English translation) Amer. Math. Soc., Providence, RI, 2001.
- [31] H. Prodinger, Compositions and Patricia tries: No fluctuations in the variance!, in: L. Arge, G. Italiano, R. Sedgewick (Eds.), *Proceedings of the Sixth Workshop on ALENEX and the First Workshop on ANALCO*, New Orleans, 2004, SIAM, Philadelphia, PA, 2004, pp. 211–215. Available from: http://math.sun.ac.za/~prodinger/pdffiles/new_orleans.pdf.
- [32] R. Sedgewick, P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Boston, 1996.
- [33] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley-Interscience, New York, 2001.
- [34] W. Vervaat, Limit theorems for records from discrete distributions, *Stochastic Processes Appl.* 1 (1973) 317–334.