**Probability Surveys**
Vol. 16 (2019) 1–61
ISSN: 1549-5787
https://doi.org/10.1214/13-PS221

# Size bias for one and all*

**Richard Arratia**

*Department of Mathematics*
*University of Southern California*
*Los Angeles CA 90089*
*e-mail:* rarratia@usc.edu

**Larry Goldstein**

*Department of Mathematics*
*University of Southern California*
*Los Angeles CA 90089*
*e-mail:* larry@usc.edu

**and**

**Fred Kochman**

*Center for Communications Research*
*805 Bunn Drive*
*Princeton, NJ 08540*
*e-mail:* kochman@idaccr.org

**Abstract:** Size bias occurs famously in waiting-time paradoxes, undesirably in sampling schemes, and unexpectedly in connection with Stein's method, tightness, analysis of the lognormal distribution, Skorohod embedding, infinite divisibility, branching processes, and number theory. In this paper we review the basics and survey some of these unexpected connections.

Received August 2013.

## Contents



*This is an original survey paper

## 1. Prologue: the waiting time paradox

In the famous "waiting time paradox", see Feller [38, Section I.4], there are two plausible but conflicting analyses of the waiting time for the next bus, once you get to the bus stop. More formally, this paradox concerns the waiting time $W_t$ for the next arrival, starting from an arbitrary instant $t$, in a standard homogeneous Poisson process with intensity parameter $\lambda = 1$: (a) The lack of memory of the exponential interarrival time suggests that $\mathbb{E}W_t$ is not sensitive to the choice of $t$; so $\mathbb{E}W_t = \mathbb{E}W_0 = 1$. (b) Since the starting time is chosen uniformly in the interval between two successive arrivals, an interval of mean length 1, symmetry suggests that $\mathbb{E}W_t = 1/2$.

As Feller shows, the reasoning behind *both* analyses is faulty, because it is the instant and not the interval which is arbitrary: a longer interval thereby becomes more likely than the relative frequencies of interarrival lengths would suggest, a

canonical instance of size biasing. So an unqualified appeal to properties of the original interarrival distribution is fallacious.

In fact, as we will discuss, a reasonable but precise interpretation of "arbitrary instant" leads to the answer given in (a), though not for the reason given in (a).

Not just recreational chestnuts, but also practical matters, such as statistical sampling tasks, are bedeviled by size bias; we provide a few references later. Surprisingly, however, size bias plays a role in such unexpected contexts as Stein's method, Skorohod embedding, nonuniqueness in the method of moments, infinite divisibility of distributions, branching processes, and number theory. We will return to the "paradox" shortly, after giving the basics of size bias. Then we will survey size bias as it appears in some of the non-sampling contexts.[1]

In [7, pp. 78–80], the authors introduce their two and one half page survey of size bias by saying "Size-biasing arises naturally in statistical sampling theory (cf. Hansen and Hurwitz (1943) [46], Midzuno (1952) [66] and Gordon (1993) [44]), and the results we present below are all well known in the folk literature." In the present paper, we feel that we have contributed a number of new results: the conceptual heuristic given in Section 3 to explain (26), where a sum of independent variables is size biased by biasing only a single term, the explanation of an intimate connection between uniform integrability and tightness in Section 8, the size bias perspective on Skorohod embedding in 10, and the treatment of infinite divisibility in Section 11 — at least the argument based on (85), size biasing a sum by size biasing a single summand.

Another survey of size bias, with a different focus, is [24].

## 2. Size bias basics

### *2.1. Bias in general*

Let $h$ be a nonnegative function, and $X$ be a random variable taking values in the domain of $h$, with $\mathbb{E}h(X) \in (0, \infty)$. For such $X$ and $h$, we say $X^h$ has the $h$-biased $X$ distribution if and only if the distribution of $X^h$, relative to the distribution of $X$, has Radon-Nikodym derivative given by

$$\frac{\mathbb{P}(X^h \in dx)}{\mathbb{P}(X \in dx)} = \frac{h(x)}{\mathbb{E}h(X)}. \tag{1}$$

The support of (the distribution of) $X^h$ is then a subset of the support of $X$, possibly a proper subset due to the set where $h$=0:

$$\mathrm{supp}(X^h) = (\mathrm{supp}(X) \setminus h^{-1}(0))^{cl}, \tag{2}$$

where $A^{cl}$ denotes the closure of $A$. A nice pair of examples, both having equal support for $X$ and $X^*$, and using $h(x) = x$, is presented in Figure 2.1 on page 15.

[1] An early draft of the present paper, with the title 'Size biasing, when is the increment independent?', has been circulated since 1998, and was cited in [72]; an update, 'Size bias, sampling, the waiting time paradox, and infinite divisibility: when is the increment independent?' was cited in [70, 71, 77]. Both of these drafts are superseded by this paper.

The class of exponential functions, $h(x) = e^{\beta x}$ for various choices of $\beta \in (-\infty, \infty)$, is very important. This class is central to exponential families and large deviation theory, but no single value $\beta$ plays a special role. The family of power functions $h(x) = x^\beta$ for $\beta > 0$ might be viewed as runner up, behind the family of exponential functions, but here the choice $\beta = 1$ is truly special. We believe that $h(x) = x$ for $x \geq 0$ is *the most important* example of bias.

### *2.2. Size bias in particular*

When $h$ is the function $h(x) = x$ with domain $[0, \infty)$, the $h$-bias above is called size bias. Thus, one can size bias the distribution of any nonnegative random variable $X$ for which $a := \mathbb{E}X \in (0, \infty)$. Instead of $X^h$ one writes $X^*$ or $X^{\mathrm{s}}$ for a random variable with the size-biased distribution of $X$. The characterization (1) reduces to

$$\frac{\mathbb{P}(X^* \in dx)}{\mathbb{P}(X \in dx)} = \frac{x}{a}. \tag{3}$$

For the common special cases, where $X$ is discrete with probability mass function $f$, or where $X$ is absolutely continuous with density $f$, the formula

$$f_{X^*}(x) = \frac{xf(x)}{a}, \tag{4}$$

completely specifies the size-biased distribution.

Does size bias commute with conditioning, on events of the form $(X \in B)$? The answer, of course, is yes — provided that $\mathbb{P}(B) > 0$. This is made obvious using the bias-in-general viewpoint of Section 2.1: any two biasings commute, because multiplication is commutative. In detail: suppose that $g$, like $h$ in Section 2.1, is a nonnegative function whose domain includes the support of $X$, and $\mathbb{E}g(X) \in (0, \infty)$. Then one can bias with respect to $g$, to specify the distribution of $X^g$. Elementary conditioning, on the event $(X \in B)$, is precisely the case where $g$ is the indicator function for $B$; in this case $\mathbb{E}g(X) = \mathbb{P}(B) < \infty$, and in the phrase *elementary conditioning*, the word elementary means that $\mathbb{P}(B) > 0$. Back to the general case: suppose that the product $gh$ (the pointwise product, not the composition $(g \circ h)(x) = g(h(x)),$) also has strictly positive, finite expectation, i.e., $\mathbb{E}(g(X)h(X)) \in (0, \infty)$. Then the iterated biased distributions, of $(X^h)^g$, and of $(X^g)^h$ are, of course, equal to each other, since they are both the same as the distribution of $X^{(gh)}$.

An interesting case of (4) involves the Poisson distributions. Starting from the assumption that the distribution of $X$ is Poisson $(\lambda)$, so that $f(k) = e^{-\lambda}\lambda^k/k!$, then (4) with $x = k + 1$ for $k = 0, 1, 2, \ldots$ gives

$$f_{X^*}(k+1) = \frac{(k+1)f(k+1)}{\lambda} = \frac{k+1}{\lambda}e^{-\lambda}\frac{\lambda^{k+1}}{(k+1)!} = f(k). \tag{5}$$

Hence for $X$ with a Poisson $(\lambda)$ distribution, $0 < \lambda < \infty$,

$$X^* =^d X + 1. \tag{6}$$

The above result is sometimes called *Robbins' Lemma*.

Conversely,

**Proposition 2.1.** *Suppose that $X$ is a nonnegative* integer-valued *random variable with mean $\lambda \in (0, \infty)$, and $X^* =^d X + 1$. Then $X$ is Poisson $(\lambda)$.*

*Proof.* Equation (4) shows that for $k \geq 0$, the point mass function $f$ for $X$ satisfies $f(k+1) = \lambda f(k)/(k+1)$, hence by induction $f(k) = f(0)\lambda^k/k!$. The assumption that $X$ is nonnegative *integer-valued* implies that $\sum_{k\geq 0} f(k) = 1$, hence $1 = \sum_{k\geq 0} f(0)\lambda^k/k! = f(0)e^{\lambda}$. □

The observation that a nonnegative integer-valued random variable $X$ is Poisson($\mathbb{E}X$) if and only if $X^* =^d X + 1$ may be viewed as the starting point for Stein's method for the Poisson distribution; see Section 5.

It is also true that if $X \geq 0$ and $0 < \lambda := \mathbb{E}X < \infty$ and $X^* =^d X + 1$, (without assuming that $X$ is *integer-valued*,) then $X$ is Poisson $(\lambda)$. This is not so obvious, and the reader might enjoy giving his own elementary proof, by combining the support consideration (2) with (4); alternately, see Theorem 11.2 and Corollary 11.3.

If $X$ is Bernoulli($p$), meaning that $\mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p$, and if $p > 0$, then $X$ can be size biased. Using either the support consideration (2), or the mass function formula (4), we see that

$$\text{for } X \sim \text{Bernoulli}(p), 0 < p \leq 1, \quad X^* = 1. \tag{7}$$

This Bernoulli family example shows that the size bias transformation is not one to one.

It is easy to see that (3) is implied by

$$\text{for all bounded } \textit{continuous } g, \quad \mathbb{E}g(X^*) = \frac{1}{a}\,\mathbb{E}(Xg(X)), \tag{8}$$

and that (3) implies[2]

$$\text{for all bounded } \textit{measurable } g, \quad \mathbb{E}g(X^*) = \frac{1}{a}\,\mathbb{E}(Xg(X)). \tag{9}$$

Even in the discrete and absolutely continuous cases, where the elementary identity (4) applies, the characterization of size bias via (9) is very handy for manipulations.

### *2.2.1. Generating functions*

Let $\phi \equiv \phi_X$ be the characteristic function of $X$, so that $\phi(u) = \mathbb{E}e^{iuX}$. A standard fact, for example from [38, XV.4 Lemma 2], is that $\mathbb{E}|X| < \infty$ implies

[2]Some would say, by definition, (3) means exactly (9), others might say that by definition, (3) means (9) restricted to $g$ being the indicator function of a measurable subset of $[0, \infty)$.

$\phi$ is differentiable, with $\phi'(u) = i\mathbb{E}(Xe^{iuX})$. With $g(x) = e^{iux}$, by taking real and imaginary parts, (9) implies that for nonnegative $X$ with $a := \mathbb{E}X \in (0, \infty)$,

$$\phi_{X^*}(u) := \mathbb{E}e^{iuX^*} = \frac{1}{a}\mathbb{E}(Xe^{iuX}) = \frac{1}{i\,a}\phi'_X(u), \tag{10}$$

and since characteristic functions determine distribution, (10) also completely specifies the size bias distribution. Suppressing the dummy variable to get a clean display, (10) says

$$\phi'_X = i\ \mathbb{E}X\ \phi_{X^*}\ .$$

In case the nonnegative random variable $X$ above is integer valued, one could use probability generating functions instead of characteristic functions to characterize the distributions of $X$ and $X^*$. With random variable name $N$ in place of $X$, and $p_n := \mathbb{P}(N = n)$, we have $G_N(z) = \sum p_n z^n$ with derivative $G'_N(z) = \sum np_n z^{n-1}$. So if $0 < \mathbb{E}N < \infty$, the generating function for the size biased random variable is

$$\begin{aligned} G_{N^*}(z) &:= \sum_{n\geq 0} \mathbb{P}(Z^* = n)z^n = \sum \frac{np_n}{\mathbb{E}N} z^n = \frac{z}{\mathbb{E}N}\ G'_N(z);\ G_{(N^*-1)}(z) \\ &= \frac{1}{\mathbb{E}N}\ G'_N(z)\ . \end{aligned}$$

Suppressing the dummy variable to get a clean display, this relation is

$$G'_N = \mathbb{E}N\ G_{(N^*-1)}\ . \tag{11}$$

### *2.2.2. Compound distributions for random sums*

Here is an application of (10). Suppose $N$ is a nonnegative integer valued random variable, with finite strictly positive mean, and $X, X_1, X_2, \ldots$ are i.i.d., independent of $N$. With $S_n := X_1 + \cdots + X_n$ and $Z = S_N = X_1 + \cdots + X_N$, the distribution of the random sum $Z$ is called a *compound distribution*, although the phrase itself is often used to refer to mixtures in general. With the notation $G$ for probability generating functions, and $\phi$ for characteristic functions, the characteristic function of $Z = X_1 + \cdots + X_N$ is $\phi_Z = G_N \circ \phi_X$. Now if $X \geq 0$ and $0 < \mathbb{E}X < \infty$, so that $X$ can be size biased, then $Z$ can also be size biased. From (10) we have

$$\phi_{Z^*}(u) = \frac{1}{i\,\mathbb{E}Z}\phi'_Z(u) = \frac{1}{i\,\mathbb{E}Z}(G_N(\phi_X(u)))'$$

so from the chain rule, and $\mathbb{E}Z = \mathbb{E}N\,\mathbb{E}X$, and (11), we have

$$\begin{aligned} \phi_{Z^*}(u) &= \frac{1}{i\,\mathbb{E}Z} \times G'_N(\phi_X(u)) \times \phi'_X(u) \\ &= \frac{1}{i\,\mathbb{E}N\,\mathbb{E}X} \times \mathbb{E}N\ G_{(N^*-1)}(\phi_X(u)) \times i\ \mathbb{E}X\ \phi_{X^*}(u) \\ &= G_{(N^*-1)}(\phi_X(u)) \times \phi_{X^*}(u). \end{aligned} \tag{12}$$

Since a product of two characteristic functions gives the distribution of the sum of two independent random variables, (12) specifies a rule:

**Rule for size biasing a random sum**. To size bias a random sum $Z = S_N$ of $N$ independent copies of $X$, where both $N$ and $X$ are nonnegative, with strictly positive finite mean: (1) Size bias $N$ to get $N^*$, and then take *one less*, $N^* - 1$, as the new compounding variable, and (2) Add in an independent copy of the size biased version of the summand, $X^*$. To say the same, less formally: size bias the number of summands, and replace one summand by a size biased version.

Of course, specializing $N$ to be concentrated at a fixed positive integer $n$ immediately yields a rule for size biasing a sum of $n$ independent identically distributed terms. However, we will rederive that rule, as (30) in Section 2.4 below, which studies how to size bias a sum of (a nonrandom number of) random variables, with no need to assume either independence or identical distribution for the summands.

### *2.2.3. Unbounded functions, and moments*

Recall that for a real valued random variable, "$\mathbb{E}Y \in [-\infty, \infty]$ exists" means that it is *not* the case that *both* the positive and the negative parts of $Y$ have infinite expectation. We extend slightly the statement that, if $\mathbb{E}|Xg(X)| < \infty$, then $\mathbb{E}g(X^*) = \mathbb{E}(Xg(X))/\mathbb{E}X$.

**Lemma 2.2.** *Let $g : [0, \infty) \to \mathbb{R}$ be measurable, and let $X$ be a nonnegative random variable with $a := \mathbb{E}X \in (0, \infty)$.*

$$\text{If } \mathbb{E}(Xg(X)) \in [-\infty, \infty] \text{ exists, then } \mathbb{E}g(X^*) = \frac{1}{a}\,\mathbb{E}(Xg(X)). \tag{13}$$

*If $\mathbb{E}(Xg(X))$ doesn't exist in $[-\infty, \infty]$, then neither does $\mathbb{E}g(X^*)$.*

*Proof.* In outline, the proof is: consider separately the positive and negative parts of $g$; for each of these, apply (9) to truncations, and apply monotone convergence.

In detail: when $g(x) \geq 0$, by applying (9) to $g_n(x) = \max(g(x), n)$, and taking limits, we conclude that

$$\mathbb{E}g(X^*) = \frac{1}{a}\,\mathbb{E}(Xg(X))$$

holds, *including* the case where both sides are infinite. Write $y_+$ and $y_-$ for the positive and negative parts of $y$. Then the functions $g_+$ and $g_-$ given by $g_+(x) = (g(x))_+$ and $g_-(x) = (g(x))_-$ are nonnegative. Note that on the domain $[0, \infty)$, $(xg(x))_+ = xg_+(x)$ and $(xg(x))_- = xg_-(x)$. Under the hypothesis that $\mathbb{E}(Xg(X)) \in [-\infty, \infty]$ exists, at least one of $h = g_+$ and $h = g_-$ has $\mathbb{E}(Xh(X)) < \infty$, and hence $\mathbb{E}g(X^*) = \mathbb{E}g_+(X^*) - \mathbb{E}g_-(X^*) \in [-\infty, \infty]$ is well defined, with value given by $(1/a)\mathbb{E}(Xg_+(X)) - (1/a)\mathbb{E}(Xg_-(X)) = (1/a)\mathbb{E}(X(g(X))$. Likewise, when $\mathbb{E}((Xg(X))_+) = \mathbb{E}((Xg(X))_-) = \infty$, we have both $\mathbb{E}g_+(X^*) = \mathbb{E}g_-(X^*) = \infty$ so that $\mathbb{E}g(X^*)$ does not exist. □

In particular, taking $g(x) = x^n$ in (13), we have

$$\mathbb{E}(X^*)^n = \mathbb{E}X^{n+1}/\mathbb{E}X \tag{14}$$

and this includes the case where both sides are infinite. Apart from the extra scaling by $1/\mathbb{E}X$, (14) says that the sequence of moments of $X^*$ is the sequence of moments of $X$, but shifted by one. Hence one way to recognize size biasing is through the *shift of the moment sequence*; this plays a role in two interesting examples, (18) and (62).

### *2.2.4. Stochastic monotonicity*

It is easy to see that, in general, $X^*$ lies above $X$ in distribution, i.e., $\mathbb{P}(X^* > t) \geq \mathbb{P}(X > t)$ for all $t$. In detail: letting $g(x) = 1(x > t)$ in (9) for some fixed $t$,

$$\mathbb{P}(X^* > t) = \frac{\mathbb{E}(X1(X > t))}{\mathbb{E}X} \geq \frac{\mathbb{E}X\ \mathbb{E}1(X > t)}{\mathbb{E}X} = \mathbb{P}(X > t) \tag{15}$$

where the inequality above is the special case $f(x) = x$, $g(x) = 1(x > t)$ of Chebyschev's correlation inequality: $\mathbb{E}(f(X)g(X)) \geq \mathbb{E}f(X)\ \mathbb{E}g(X)$ for any random variable and any two increasing functions $f, g$.

The condition $\mathbb{P}(X^* > t) \geq \mathbb{P}(X > t)$ for all $t$ is described as "$X^*$ lies above $X$ in distribution," written $X \leq_{st} X^*$, and implies that there exist couplings of $X^*$ and $X$ in which always $X \leq X^*$. Writing $Y$ for the difference, we have

$$X^* = X + Y, \quad Y \geq 0. \tag{16}$$

In general, the known marginals for $X$ and $X^*$ do not uniquely determine the distribution of a coupling; in Section 11 we will study the question: when can (16) be achieved with $X, Y$ independent?

Suppose the distribution of $Z$ is defined to be that of $X$, conditional on $(X > 0)$. Recalling the third paragraph of Section 2.2, it is obvious that $X^* =^d Z^*$. And of course, $Z$ lies above $X$ in distribution since for $t \geq 0$, $\mathbb{P}(X > t|X > 0) = \mathbb{P}(X > t)/\mathbb{P}(X > 0) \geq \mathbb{P}(X > t)$. To summarize, for nonnegative $X$ with $\mathbb{E}X \in (0, \infty)$, we have the stochastic monotonicity sandwich

$$X \leq_{st} (X|X > 0) \leq_{st} X^*.$$

### *2.2.5. Scaling, coupling, and limits in distribution*

It is easy to see, from (9), that size biasing respects multiplication by positive constants, that is, with $c > 0$,

$$(cX)^* =^d c(X^*). \tag{17}$$

The notation used above, $X =^d Y$, is often written $\mathcal{L}(X) = \mathcal{L}(Y)$, to say that random variables $X$ and $Y$ have the same law, or distribution. The simpler notation $X = Y$ would imply a coupling, i.e., that $X$ and $Y$ are defined on the same probability space, with $X(\omega) = Y(\omega)$ for all outcomes $\omega$.

It is also true that size bias respects convergence in distribution, provided one is careful to make the additional hypothesis that the means converge to the mean of the limit random variable, which is in this context equivalent to uniform integrability.

**Theorem 2.3.** *Suppose that $X, X_1, X_2, \ldots$ are nonnegative random variables with $a := \mathbb{E}X \in (0,\infty)$, $a_n := \mathbb{E}X_n \in (0,\infty)$, that $X_n \Rightarrow X$, and that $a_n \to a$. Then*

$$X_n^* \Rightarrow X^*.$$

*Proof.* Let $h : \mathbb{R} \to \mathbb{R}$ be a bounded continuous function with compact support. Then the function $g$ given by $g(x) = x\,h(x)$ is bounded and continuous. Since $g$ is bounded, (9) applies, and since $g$ is continuous, the hypothesized distributional convergence implies $\mathbb{E}g(X_n) \to \mathbb{E}g(X)$. Using (9) with $h$ in the role of $g$, we have

$$\mathbb{E}h(X_n^*) = \frac{\mathbb{E}X_n h(X_n)}{a_n} = \frac{\mathbb{E}g(X_n)}{a_n} \to \frac{\mathbb{E}g(X)}{a} = \frac{\mathbb{E}Xh(X)}{a} = \mathbb{E}h(X^*). \qquad \square$$

The necessity of the hypothesis that $\mathbb{E}X > 0$, in Theorem 2.3, is shown by the example with $X_n$ distributed as Bernoulli$(1/n)$, so that $X_n^* \Rightarrow 1$, and $X_n \Rightarrow X = 0$, but the limit random variable $X$ cannot be size biased.

The converse of Theorem 2.3 is false, since the correspondence $\mathcal{L}(X) \mapsto \mathcal{L}(X^*)$ is many to one. In detail, take any $A, B$ with $A \neq^d B$ and $A^* =^d B^*$; then the sequence $X_1, X_2, X_3, X_4, \ldots = A, B, A, B, \ldots$, together with $X = A$, has $X_n^* \Rightarrow X^*$ but not $X_n \Rightarrow X$.

An interesting natural example, related to the non-converse of Theorem 2.3, involves $X_n$ which cannot be rescaled to have a nontrivial limit distribution, while the corresponding $X_n^*$ can. Take $X_n$ to have the Borel distribution[3] with parameter $\lambda = 1 - 1/n$. Calculation shows that $\mathbb{E}(X_n) = n$ and for $k = 1, 2, \ldots$, $\mathbb{E}(X_n)^{k+1} \sim n^{2k+1}(2k-1)(2k-3)\cdots 5 \times 3 \times 1$, hence one cannot scale the $X_n$ sequence to get a nontrivial distributional limit. But using (14), we have

$$\mathbb{E}(X_n^*)^k \sim n^{2k}(2k-1)(2k-3)\cdots 5 \times 3 \times 1, \tag{18}$$

so that with $Z$ for a standard normal, $X_n^*/n^2 \Rightarrow Z^2$.

### *2.2.6. Mixtures, biasing a conditional probability*

First, we give Lemma 2.4, an elementary result on how to size bias a mixture of distributions. An application of Lemma 2.4 will be given by Lemma 9.2 and the subsequent Theorem 9.4. Mixtures are often discussed in conjunction with regular conditional probabilities; see for example [28, 61].

Suppose that $I \subset \mathbb{R}$, that $h$ is a probability measure on $I$, that for each $b \in I$ $\mu_b$ is a distribution for a nonnegative random variable $X_b$, with $m(b) := \mathbb{E}X_b \in (0,\infty)$, and that $b \mapsto \mu_b$ is measurable. Note, we have assumed that for every $b$, $m(b) \in (0,\infty)$ in order that, for every $b$, the size-biased distribution for $X_b^*$ be

[3] For $\lambda \in [0,1)$, one says $X$ has the Borel$(\lambda)$ distribution if $X$ is the total progeny in the subcritical Galton-Watson branching process where the individual offspring distribution is Poisson with mean $\lambda$, equivalently, $\mathbb{P}(X = i) = \exp(-\lambda\, i)(\lambda\, i)^{i-1}/i!$ for $i = 1, 2, \ldots$; see [2].

defined. We say that (the distribution of) $X$ is the mixture of (the distributions of) $X_b$, governed by $h$, if for all bounded measurable $g$,

$$\mathbb{E}g(X) = \int \mathbb{E}g(X_b)\ dh(b).$$

Of course, for such a mixture, $\mathbb{E}X = \int m(b)\ dh(b) \in (0, \infty]$, but since we are interested is size bias, we make the additional assumption that $a := \mathbb{E}X < \infty$.

**Lemma 2.4.** *Under the setup of the previous paragraph, with $a = \int m(b)\, dh(b) \in (0, \infty)$, the distribution of $X^*$ is a mixture of the distributions of the $X_b^*$. The measure $h^s$ governing this mixture is defined in terms of the original governor $h$ via its Radon-Nikodym derivative, $dh^s(b)/dh(b) = m(b)/a$. In particular, if $m(b)$ is constant, then $h^s = h$, i.e., the measure governing $X^*$ as a mixture of the $X_b^*$ is equal to the measure governing $X$ as a mixture of the $X_b$.*

*Proof.* For bounded measurable $g$

$$\mathbb{E}g(X^*) = \frac{\mathbb{E}(Xg(X))}{a} = \int \frac{\mathbb{E}(X_b g(X_b))}{m(b)}\ \frac{m(b)\ dh(b)}{a} = \int \mathbb{E}g(X_b^*)\ dh^s(b). \quad \square$$

In a different direction, the following result from [41] can be useful for constructing size bias couplings for continuous random variables that are not represented as sums, though it may also be noted that Lemma 2.5 implies (26) for sums of indicator variables, see [16, Lemma 2.6 ff].

**Lemma 2.5.** *Let $X = \Pr(A|\mathcal{F})$ where $\mathcal{F}$ is some $\sigma$-algebra and $A$ is some event with $0 < \Pr(A) < 1$. Then $X^*$ has the distribution of $X$ conditioned on $A$.*

*Proof.* For any bounded measurable $g$, we have

$$\mathbb{E}g(X^*) = \frac{\mathbb{E}(g(X)\mathbb{E}(1_A|\mathcal{F}))}{\mathbb{E}X} = \frac{\mathbb{E}(\mathbb{E}(g(X)1_A|\mathcal{F}))}{\mathbb{P}(A)}$$

$$= \frac{\mathbb{E}(g(X)1_A)}{\mathbb{P}(A)} = \mathbb{E}[g(X)|A]. \quad \square$$

### *2.2.7. Many to one, one to one*

We describe the preimage, under size biasing, of a random variable $Z$. Note first that if $Z =^d X^*$, then for any mixture $\mathcal{M} = b\delta_0 + (1-b)\mathcal{L}(X)$ with $0 \le b < 1$, a random variable $Y$ with $\mathcal{L}(Y) = \mathcal{M}$ is also a preimage. We claim that changing the amount of point-mass at 0 is the only source of non-uniqueness.

**Lemma 2.6.** *A random variable $Z$ satisfies $Z =^d X^*$ for some $X$ iff $1 = \mathbb{P}(Z > 0)$ and $\mathbb{E}(1/Z) < \infty$, and then there is a unique law for $Y > 0$ such that any $X$ having $X^* =^d Z$ is distributed as $b\delta_0 + (1-b)\mathcal{L}(Y)$ for some $0 \le b < 1$.*

*Proof.* Let $Z =^d X^*$ for some $X$; this implies $X \geq 0$, $0 < \mathbb{E}X < \infty$, and $\mathbb{P}(Z > 0) = 1$. Let $b := \mathbb{P}(X = 0)$, so clearly $b \in [0, 1)$. Let $Y$ have the distribution of $X$ conditioned on $X > 0$, so $Y > 0$, $Z =^d Y^*$, and $\mathcal{L}(X) = b\delta_0 + (1 - b)\mathcal{L}(Y)$. With $c = \mathbb{E}X/(1 - b) = \mathbb{E}Y \in (0, \infty)$, we have, as in (3), that the distributions $\nu$ of $Z$ and $\mu$ of $Y$, as measures on $(0, \infty)$, are mutually absolutely continuous, with Radon-Nikodym derivative

$$\frac{\nu(dx)}{\mu(dx)} \equiv \frac{\mathbb{P}(Z \in dx)}{\mathbb{P}(Y \in dx)} = \frac{x}{c}.$$

This shows the uniqueness of the law for $Y$; that $\mathbb{E}(1/Z) < \infty$ follows from the explicit calculation

$$\mathbb{E}\frac{1}{Z} = \int_{0<x<\infty} \frac{1}{x}\ \nu(dx) = \int \frac{1}{x}\ \frac{d\nu}{d\mu}\ \mu(dx) = \int \frac{1}{c}\ \mu(dx) = \frac{1}{c}.$$

Conversely, if $Z > 0$ with probability measure $\nu(dz)$ satisfies $0 < \mathbb{E}(1/Z) < \infty$, then with $1/c = \mathbb{E}(1/Z)$, the law $\mu$ on $(0, \infty)$ with $\mu(dy)/\nu(dy) = c/y$, as the distribution for $Y$, yields $Z =^d Y^*$.

□

A paraphrase of Lemma 2.6 is that size bias is a bijection, between equivalence classes of distributions for nonnegative random variables with strictly positive finite mean, modulo varying the size of the point mass at zero, and distributions for strictly positive random variables having finite minus first moment.

### *2.3. To bias a process by one coordinate*

The following is taken from [43]. Readers who dislike technicalities might prefer to jump directly to Section 2.4, which leads up to (23), and then come back only if they feel uncomfortable that our proof of (27) doesn't involve any limits! Suppose that $\mathbf{X} = (X_1, X_2, \ldots) \in [0, \infty)^{\mathbb{N}}$ has joint law $\mu$, and for a particular choice of $i$, $a_i := \mathbb{E}X_i \in (0, \infty)$. To bias by $X_i$ means, analogous to (3), to switch to the joint law $\mu^{(i)}$ on $[0, \infty)^{\mathbb{N}}$ with Radon-Nikodym derivative

$$\frac{d\mu^{(i)}}{d\mu} = \frac{x_i}{a_i}. \tag{19}$$

We write $\mathbf{X}^{(i)} = (X_1^{(i)}, X_2^{(i)}, \ldots)$ for a process having this joint distribution $\mu^{(i)}$. Equivalent to (19) is the following statement,

$$\text{for all bounded measurable } g, \quad \mathbb{E}g(\mathbf{X}^{(i)}) = \frac{1}{a_i}\ \mathbb{E}(X_i g(\mathbf{X})), \tag{20}$$

which looks very much like (9), except that now we have $g : [0, \infty)^{\mathbb{N}} \to \mathbb{R}$. Note that given a bounded measurable $h : [0, \infty) \to \mathbb{R}$, applying (20) to the special case $g(\mathbf{x}) := h(x_i)$ shows that our notion of process bias by one coordinate, restricted to viewing that coordinate, agrees with the original notion of size

bias, i.e., $X_i^{(i)} =^d X_i^*$. In general, there is no similarly compact description of what happens to the other coordinates. However, as we will see in Section 7.1, if the process $\mathbf{X}$ is a martingale then biasing the process by any single coordinate results in size-biasing the marginal distribution of *each* coordinate simultaneously.

In a different direction, suppose that under $\mu$ the coordinates are initially independent. Then as we now show, after biasing by the $i^{\text{th}}$ coordinate they remain independent, and only the $i$th coordinate is affected.

**Lemma 2.7.** *Fix a particular value $i$. Assume that $X_1, X_2, \ldots$ are mutually independent, nonnegative, and that $0 < \mathbb{E}X_i < \infty$. For $j \neq i$ let $Y_j =^d X_j$, let $Y_i =^d X_i^*$, and let $Y_1, Y_2, \ldots$ be mutually independent. Then the law $\mu^{(i)}$ for $\mathbf{X}^{(i)}$, as given by (19), reduces to the law for $\mathbf{Y} = (Y_1, Y_2, \ldots)$, i.e.*

$$(X_1^{(i)}, X_2^{(i)}, \ldots) =^d (Y_1, Y_2, \ldots).$$

*Proof.* First we check that the marginals match, i.e., that for each $j$, $X_j^{(i)} =^d Y_j$. We already noted that this is so, for $j = i$, as a consequence of (20), even without the hypothesis of mutual independence. For $j \neq i$, and a bounded measurable $h : [0, \infty) \to \mathbb{R}$, applying (20) to the special case $g(\mathbf{x}) := h(x_j)$ yields the relation $\mathbb{E}g(\mathbf{X}^{(i)}) = \mathbb{E}h(X_j^{(i)}) = (1/a_i)\mathbb{E}(X_i h(X_j))$. Using the independence of $X_i$ and $X_j$, we get $\mathbb{E}h(X_j^{(i)}) = (1/a_i)\mathbb{E}(X_i h(X_j)) = (1/a_i)(\mathbb{E}X_i)\mathbb{E}h(X_j) = \mathbb{E}h(X_j)$, proving that for $j \neq i$, $X_j^{(i)} =^d X_j$, as required, since for $j \neq i$, $Y_j =^d X_j$.

Next we show that $\mathbf{X}^{(i)}$ and $\mathbf{Y}$ have the same joint distribution, either by showing that $\mathbf{X}^{(i)}$ has independent coordinates, or by checking that for all measurable $C \subset [0, \infty)^{\mathbb{N}}$, $\mathbb{P}(\mathbf{X}^{(i)} \in C) = \mathbb{P}(\mathbf{Y} \in C)$, first by checking finite-dimensional cylinder sets, then applying the $\pi - \lambda$ theorem — either route seems to require the same work. Without loss of generality, the cylinder set $C$ includes a restriction on the $i^{\text{th}}$ coordinate, i.e., it has the form $C = (X_i \in B_i) \cap \bigcap_{j \in J}(X_j \in B_j)$, where $i \notin J$. Write $g_1(\mathbf{x}) = 1(x_i \in B_i)$ and $g_2(\mathbf{x}) = 1(x_j \in B_j \text{ for } j \in J)$. With $g = g_1 g_2$ in (20), calculation that $\mathbb{E}g(\mathbf{X}^{(i)}) = \mathbb{E}g(\mathbf{Y})$ is a simple extension of the calculation for the special case where the cylinder restricts only one coordinate, given in the first paragraph of this proof. □

Another technical issue involves the value infinity. It would have been possible to present the basic discussion of size bias, in particular (3) and (9), in terms of a random element $Y$ with values in $[0, \infty]$. But since $0 < \mathbb{E}Y < \infty$ implies $\mathbb{P}(Y = \infty) = 0$, it is of course possible, and simpler, to deal with $Y$ taking values in $[0, \infty)$, and this is what everyone does. However, in dealing with infinite sums of finite nonnegative random variables, one cannot simply declare that the space of values for the sum be taken as $[0, \infty)$, even if one knows that the sum is finite with probability one.

Our goal is to deal with the distribution of random variables $Y = h(\mathbf{X})$, such as $Y = X_1 + X_2 + \cdots$,[4] and to specify the distribution of $Y^{(i)}$, distributed as

[4]Thanks to only having nonnegative numbers for the coordinates of the domain, there are

$Y$ with $\mu$ changed to $\mu^{(i)}$. Hence we consider measurable $h : [0,\infty)^{\mathbb{N}} \to [0,\infty]$, and bounded measurable $f : [0,\infty] \to \mathbb{R}$. The composition $g(\mathbf{X}) = f(h(\mathbf{X}))$ is a bounded measurable function from $[0,\infty)^{\mathbb{N}} \to \mathbb{R}$, hence (20) applies. The distribution of $Y^{(i)}$ is then specified by

$$\text{for bounded measurable } f : [0,\infty] \to \mathbb{R}, \quad \mathbb{E}(f(Y^{(i)})) = \frac{1}{a_i}\,\mathbb{E}(X_i f(Y)). \tag{21}$$

### *2.4. To size bias a sum*

Consider a finite sum $S = X_1 + \cdots + X_n$, $n \geq 1$, or an infinite sum $S = X_1 + X_2 + \cdots$, with $X_i \geq 0$ and $a_i := \mathbb{E}X_i > 0$, and $a = \mathbb{E}S < \infty$. After biasing by $X_i$, as in (19), we have a sum[5] $S^{(i)} = X_1^{(i)} + \cdots + X_n^{(i)}$, so that, as a special case of (21), for bounded nonnegative measurable $g$,

$$\mathbb{E}g(S^{(i)}) = \frac{1}{a_i}\mathbb{E}(X_i g(S)),$$

and then with (9) to justify the first line, and elementary algebra (here using $g \geq 0$) to justify the second line,

$$\begin{aligned} \mathbb{E}g(S^*) &= \mathbb{E}(Sg(S))/a \\ &= \sum_i \frac{1}{a}\,\mathbb{E}(X_i g(S)) \\ &= \sum_i \frac{a_i}{a}\,\mathbb{E}g(S^{(i)}). \end{aligned} \tag{22}$$

Suppose furthermore that the summands $X_1, X_2, \ldots$ are independent. If size biased random variables $X_1^*, X_2^*, \ldots$ are realized on the same probability space, with $(X_1, X_1^*), (X_2, X_2^*), \ldots$ mutually independent, then for each $i$, by Lemma 2.7, $S^{(i)} =^d S - X_i + X_i^*$ so that (22) simplifies to: for bounded nonnegative measurable $g$,

$$\mathbb{E}g(S^*) = \sum \frac{a_i}{a}\,\mathbb{E}g(S - X_i + X_i^*). \tag{23}$$

The result above says precisely that $S^*$ can be represented by the mixture of the distributions of $S + X_i^* - X_i$ with mixture probabilities $a_i/a$. With a random $I$ having distribution defined by

$$\mathbb{P}(I = i) = a_i/a, \tag{24}$$

and all of $I, (X_1, X_1^*), (X_2, X_2^*), \ldots$ mutually independent, the mixture formula (23) can be restated as

$$S^* =^d S - X_I + X_I^*. \tag{25}$$

no convergence issues in dealing with the sum $X_1 + X_2 + \cdots \in [0,\infty]$.

[5]Warning: our notation here conflicts with some standard expositions of Stein's method, such as [25], [15, Theorem B.1], and [45], where notation $V_i$ refers to the sum, with $i$th term omitted, size biased by the $i$th term.

In the preceding coupling, for each $i$, marginal distributions of $X_i, X_i^*$ are specified, but the joint distribution of $(X_i, X_i^*)$ is otherwise arbitrary. Allowing such dependence is important for use with Stein's method; see Section 5. Of course, mutual independence for $I, X_1, X_2, \ldots, X_1^*, X_2^*, \ldots$ implies mutual independence for $I, (X_1, X_1^*), (X_2, X_2^*), \ldots$ .

For each case, $S = X_1 + \cdots + X_n$ or $S = X_1 + X_2 + \cdots$, (25) can be written out with notation to emphasize that a single term has been biased[6]:

$$(X_1 + X_2 + \cdots + X_n)^* =^d X_1 + \cdots + X_{I-1} + X_I^* + X_{I+1} + \cdots + X_n, \quad (26)$$

and

$$(X_1 + X_2 + \cdots)^* =^d X_1 + \cdots + X_{I-1} + X_I^* + X_{I+1} + \cdots . \quad (27)$$

It is a natural abuse of notation to view (26) as a special case of (27). The reason that this is abuse, rather than the special case $X_{n+1} = X_{n+2} = \cdots = 0$ is that the identically zero random variable $X$ cannot be size biased. Specifically, $X = 0$ doesn't satisfy the conditions of the definition in (3), and size biasing this $X$, if allowed, would abrogate Lemma 2.6. Nonetheless, it is customary to follow the notational abuse that if $X = 0$ then $X^* =^d X = 0$, so that one can view (26) as the special case of (27), and later, write formulas such as (33) for a sum with infinitely many terms, without writing out a second instance for a sum with finitely many terms.

In contrast to a sum of independent nonnegative summands, which is size biased by biasing a *single* term, a product $W = X_1 X_2 \cdots X_n$, of independent, nonnegative random variables $X_1, \ldots, X_n$, each with finite, strictly positive mean, is size biased by biasing *every* factor: taking $X_1^*, \ldots, X_n^*$ independent, one has

$$W^* =^d X_1^* \cdots X_n^*. \quad (28)$$

Here, we leave the proof as an exercise; this result comes from [63]. For the case of dependent summands, the decomposition (22) is useful; in contrast, for dependent factors, we don't know of any useful relation.

An interesting example of the use of (27) involves $S = \sum_{i \geq 1} 2B_i/3^i$ with independent $B_i$, with $\mathbb{P}(B_i = 0) = \mathbb{P}(B_i = 1) = 1/2$. The cumulative distribution of this sum $S$ is known as the Cantor function; the distribution of $S$ is, by all reasonable interpretations, the uniform distribution on the Cantor middle thirds set. By (24), the random index $I$ has the geometric distribution $\mathbb{P}(I = i) = 2/3^i$ for $i = 1, 2, \ldots$, and by (7), the size biased version of $B_i$ is $B_i^* = 1 = B_i + (1 - B_i)$, so that (25) simplifies to

$$S^* =^d S + 2(1 - B_I)/3^I.$$

A closely related example, using the same $B_i$, is the standard uniform (0,1) random variable $U = \sum_{i \geq 1} B_i/2^i$. With a random index $J$ having geometric distribution $\mathbb{P}(J = i) = 1/2^i$ for $i = 1, 2, \ldots$, independent of $B_1, B_2, \ldots$, (25) simplifies to

$$U^* =^d U + (1 - B_J)/2^J = \frac{B_1}{2} + \frac{B_2}{4} + \cdots + \frac{B_{J-1}}{2^{J-1}} + \frac{1}{2^J} + \frac{B_{J+1}}{2^{J+1}} + \cdots . \quad (29)$$

[6] and hence the title of this paper

Of course, it is easy to calculate that the density of $U^*$ is $2x$ on (0,1), using (4): multiply the density of the uniform by $x$ and divide by $\mathbb{E}U = 1/2$. But perhaps the following exercise is not easy.

**Exercise** Prove, without using size bias, that the sum on the right side of (29) has density $f(x) = 2x$ on (0,1).

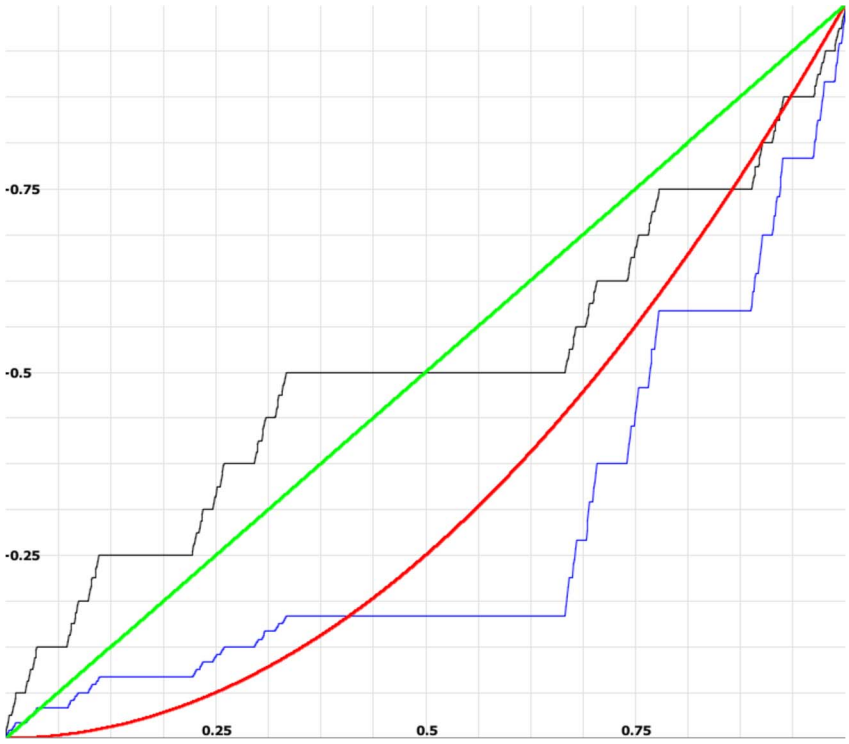


FIG 2.1. *Cumulative distribution functions for the uniform distribution on (0,1), the uniform distribution on the Cantor set, and the size biased versions of these. Image produced using MathStudio [73].*

For the case with a finite number of summands, where the summands are not only independent but also *identically distributed*, the recipe (26) simplifies. In this case it does not matter which summand is biased, as all the distributions in the mixture are the same; hence we may replace the random $I$ with the fixed $i = 1$, yielding

$$(X_1 + X_2 + \cdots + X_n)^* =^d X_1^* + X_2 + X_3 + \cdots + X_n. \tag{30}$$

Here are some elementary consequences of (30). Recall (7), that for $p \in (0, 1]$, a Bernoulli random variable with mean $p$, size biased, is the constant 1. Summing $n$ independent copies gives us random variables $S_n$ whose distribution is Binomial$(n, p)$. Hence using (30),

$$S_n^* =^d 1 + S_{n-1}. \tag{31}$$

Finally, taking $\lambda \in (0, \infty)$ fixed, $Z$ to be Poisson($\lambda$), and $X_n$ to be Binomial($n, \lambda/n$), the Poisson limit for the Binomial, together with Theorem 2.3 and (31), implies $Z^* =^d Z + 1$. Of course, this equality was already verified by direct calculation using (4) and (5), but the beauty of the argument via (31) is that it is purely conceptual.

### *2.4.1. Example: compound Poisson*

Given the distribution for a discrete positive random variable $Y$ with finite mean, and $0 < a < \infty$, we will show how to construct a distribution for $S$ such that

$$S^* =^d S + Y \text{ with } S, Y \text{ independent, and } \mathbb{E}S = a. \tag{32}$$

To specify the distribution of $Y$, suppose that $p_i = \mathbb{P}(Y = y_i)$, for distinct constants $y_1, y_2, \ldots > 0$, with $p_1 + p_2 + \cdots = 1$. The requirement $\mathbb{E}Y < \infty$ becomes $\sum p_i y_i < \infty$. Define

$$\lambda_i = a\, p_i / y_i.$$

Let $Z_i$ be Poisson with mean $\lambda_i$ with $Z_1, Z_2, \ldots$ mutually independent. We will show that

$$S = \sum_{i \geq 1} X_i, \qquad \text{with } X_i = y_i Z_i \tag{33}$$

gives a solution to (32), using only formula (6) for size biasing a single Poisson distributed random variable, the scaling property (17), formula (27) for size biasing a sum of independent, non identically distributed summands, and the trivial calculation that $a_i := \mathbb{E}X_i = \lambda_i\, y_i = a p_i$, hence $\mathbb{E}S = \sum \lambda_i\, y_i = \sum a p_i = a$.

First, using (6), $Z_i^* =^d Z_i + 1$. Second, using the scaling property (17), $X_i^* =^d X_i + y_i$. In the recipe (27), there is a random index $I$, independent of the $X_1, X_2, \ldots$, with

$$\mathbb{P}(I = i) = \mathbb{E}X_i / a = \lambda_i y_i / a = p_i, \tag{34}$$

and we can take the coupling in which $X_i^* = X_i + y_i$ for each $i$. This yields $S^* =^d S + y_I$, with $S, I$ independent. Since the $y_i$ are distinct, for each $i$, as events, $(Y_I = y_i) = (I = i)$, hence the distribution of $y_I$ *is* the given distribution for $Y$. To summarize, we were given the distribution for $Y$, and we constructed a distribution for $S$ so that (32) holds. We will revisit the relation $S^* =^d S + Y$ with $S, Y$ independent in Section 11; the preceding is then seen as an explicit example of (83), with the distribution of $Y$ specified in advance. In the standard literature, the random variable $S$ in (33) is said to have a *compound Poisson* distribution, given the further restriction that $\sum_i \lambda_i < \infty$. Compound Poisson with finite mean requires *both* $\sum_i \lambda_i < \infty$ and $\sum \lambda_i\, y_i < \infty$; in contrast, we require *only* the latter.

Recall, if $Z$ is Poisson($\lambda$) then its probability generating function is $G_Z(s) := \mathbb{E}s^Z = \exp(\lambda(s-1))$. Substituting $s = e^\beta$, the moment generating function of

$Z$ is $M_Z(\beta) := \mathbb{E}e^{\beta Z} = \exp(\lambda(e^\beta - 1))$. Hence in (33), the moment generating function of $X_i$ is $M_{X_i}(\beta) = \exp(\lambda_i(e^{\beta y_i} - 1))$ and the moment generating function of $S$ is

$$M_S(\beta) = \exp\left(\sum_i \lambda_i(e^{\beta y_i} - 1)\right) = \exp\left(a \sum_i \frac{e^{\beta y_i} - 1}{y_i} \mathbb{P}(I = i)\right), \qquad (35)$$

with the distribution of $I$ given by (34). Likewise, the characteristic function of $S$, $\phi_S(u) := \mathbb{E}e^{iuS}$ is given by

$$\phi_S(u) = \exp\left(\sum_k \lambda_k(e^{iu y_k} - 1)\right) = \exp\left(a \sum_k \frac{e^{iu y_k} - 1}{y_k} \mathbb{P}(I = k)\right). \qquad (36)$$

## 3. Waiting time paradox: the renewal theory connection

We resolve the waiting time paradox from Section 1 in the general context of renewal processes, at the same time providing a conceptual explanation of the identities (26) and (30).

Let the interarrival times in Section 1 be denoted $X_i$ so that, starting from 0, arrivals occur at times $X_1, X_1 + X_2, X_1 + X_2 + X_3, \ldots$, and assume only that the $X_i$ are i.i.d., strictly positive random variables with finite mean; the paradox presented earlier was for the special case with $X_i$ exponentially distributed.

The following argument is heuristic. One way to model the "arbitrary instant $t$" is to choose a random $T$ uniformly from 0 to $l$, independent of $X_1, X_2, \ldots$, and then take the limit as $l \to \infty$. For large but finite $l$, conditional on $X_1, X_2, \ldots$, apart from possible cutoff at the extreme right[7] the probability of $T$ landing in a given interarrival interval is proportional to its length. In other words, if the interarrival times $X_i$ have a distribution $dF(x)$, the distribution of the length of the selected interval is approximately proportional to $x\ dF(x)$. In the limit, it is precisely correct that the distribution of the length of the selected interval is the distribution of $X^*$.

For the particular case of exponentially distributed interarrival times, the density of $X^*$ is $xe^{-x}$, with mean value 2, and so a right–left symmetry argument gives the answer in a).

A conceptual explanation of identity (30) is given by the following heuristic. Group the interarrival intervals into successive blocks of $n$ intervals. By considering only the endpoints of blocks, i.e., the renewal process, decimated by $n$, the random time $T$ must find itself in a block with total length distributed as $S^* = (X_1 + \cdots + X_n)^*$. But regardless of the grouping, the random time $T$ still finds itself in an internal interval whose length is distributed as the size biased distribution of the interarrival times; the lengths of the other intervals in the same block are not affected. Thus the total block length must also be

[7] Conditional on $T = t$ and $X_1 + \cdots + X_{m-1} < t < X_1 + \cdots + X_{m-1} + X_m$, there are $m$ interarrival intervals, and for $i = 1$ to $m-1$ interval $i$ is selected with probability proportional to $X_i$, but interval $m$ is selected with probability proportional to $t - (X_1 + \cdots + X_{m-1}) < X_m$.

distributed as $X_1 + \cdots + X_{i-1} + X_i^* + X_{i+1} + \cdots + X_n$. A small extension of this heuristic may convince one of the identity (26): given $n$ distributions for strictly positive $X_1, \ldots, X_n$ all with finite mean, create the $n$-alternating renewal process, in which the independent interarrival time distributions cycle through the $n$ given distributions. The decimation by $n$ has independent interarrival times distributed as $S = X_1 + \cdots + X_n$, with independent summands, $T$ picks out a block with length distributed as $S^*$, and the contribution $\mathbb{E}X_i$ makes to the total block size governs the distribution of which subinterval in a block gets chosen by $T$. And for (27), where $S = X_1 + X_2 + \cdots$ with $\mathbb{E}S < \infty$, another small extension of the heuristic may be convincing. But we don't really expect the $\infty$-alternating renewal process to become a popular model.

The standard rigorous analysis of the waiting time paradox, for instance in [88], is a bit less direct, based on randomizing the starting point of the arrivals, so that the arrival times form a stationary sequence. Begin by extending $X_1, X_2, \ldots$ to an independent, identically distributed sequence $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$. Informally, if the arbitrary instant $t$ could be uniform on the whole line (or by adapting the above limiting argument) then $t$ would fall uniformly inside a size biased interarrival interval; relabeling, we call $t$ by the name zero, and the landing interval has length $X_0^*$. Then the prior arrival and next arrival would be at times $-(1-U)X_0^*$ and $UX_0^*$ respectively, where the uniform $U \in [0,1]$ is independent of the $X_i$'s. Thus motivated, we define a process by setting arrivals at positive times $UX_0^*, UX_0^* + X_1, UX_0^* + X_1 + X_2, \ldots$, as well as negative times $-(1-U)X_0^*, -((1-U)X_0^* + X_{-1}), -((1-U)X_0^* + X_{-1} + X_{-2}), \ldots$. It can be proved that this process is stationary, see [88, Theorem 8.1, Chapter 8]. Our desired waiting time $W_t$ is then equal in distribution to $W_0 = UX_0^*$.

The interval which covers the origin has expected length $\mathbb{E}X_0^* = \mathbb{E}X_0^2/\mathbb{E}X_0$ (by (14) with $n = 1$,) and the ratio of this to $\mathbb{E}X_0$ is $\mathbb{E}X_0^*/\mathbb{E}X_0 = \mathbb{E}X_0^2/(\mathbb{E}X_0)^2$. By Cauchy-Schwarz, this ratio is at least 1, (see also (15),) and every value in $[1, \infty]$ is feasible. Since the mean waiting time is $\mathbb{E}W_t = \mathbb{E}W_0 = \mathbb{E}(UX_0^*) = (1/2)\mathbb{E}X_0^*$, the ratio $\mathbb{E}W_t/\mathbb{E}X_0$ can be any value between $1/2$ and infinity, depending on the distribution of $X_0$.

The exponential case is very special, where "coincidences" effectively hide all the structure involved in size biasing. As suggested by Feller's argument (a) at the start of this paper, but now justified by stationarity, $\mathbb{E}W_t = 1$. Furthermore, for the exponential case, where $X_0$ has density $e^{-x}$ for $x > 0$, one gets $X_0^*$ has density $xe^{-x}$ and the two summands $UX_0^*$ and $(1-U)X_0^*$ are independent, each with the original exponential distribution.[8] Thus the general recipe for cooking up a stationary process, involving $X_0^*$ and $U$ in general, simplifies beyond recognition: the original simple process with arrivals at times $X_1, X_1 + X_2, X_1 + X_2 + X_3, \ldots$ forms half of a stationary process, which is completed by its other half, arrivals at $-X_1', -(X_1' + X_2'), \ldots$, with $X_1, X_2, \ldots, X_1', X_2', \ldots$ all independent and exponentially distributed.

The above material deals with *renewal* processes, and perhaps originated in

[8] Exercise for the reader: prove that if $U\,X^* =^d X$ when $U$ is independent of $X^*$ and $U$ is distributed uniformly on $(0,1)$, then $X$ has an exponential distribution — on some scale. Not hard; or, see [69].

Doob [36]. A broad generalization, applying to *stationary* point processes — dropping the requirement that the interarrival times be independent — was given by [55]. See also [34, p. 299].

## 4. Size bias in statistics

We now touch briefly on the topic of inadvertent or unavoidable size bias[9] in statistical sampling by citing two references from a vast literature.[10] We also discuss the *deliberate* use of size bias, as a sampling tool.

### *4.1. Inadvertent size bias*

In a 1969 paper [33] David Cox identifies, among other topics, length bias in a then-standard process for estimating the mean length of textile fibers: In outline, as he describes it, fibers are gripped by a pincer, all ungripped fibers adhering to the gripped ones are carefully removed, and the remaining fibers are measured. Cox points out that since shorter fibers are more likely to be missed by the pincer, the distribution of the sampled lengths is length biased. He proposes some adapted estimators for getting at parameters of the original distribution if the sampling process itself cannot be refined.

Nearer to the present, the 2009 paper [50] considers issues arising in assessing the value of medical screening and the effects of subsequent early treatment on survival time. As discussed in [50], for reasons analogous to waiting-time bias, the durations of preclinical disease states detected by certain screening protocols are subject to length bias. Even though the durations themselves are not observed, longer durations are likely to derive from slower-acting instances of the disease under consideration, and hence are correlated a priori with longer survival times. Therefore, as indicated by the authors, improvement in survival time is likely to be overestimated by such studies if suitable adjustments are not made.

### *4.2. Deliberate size bias to create something unbiased*

Somewhat paradoxically, size biasing can occasionally be used to construct *unbiased* estimators of quantities that would seem, at first glance, difficult to estimate without bias. The following procedure for unbiased ratio estimation is due to Midzuno [66]; see also Cochran [31]. Suppose that for each individual $i$ in some large population there is a pair of numbers $(x_i, y_i)$, with the value $x_i$ easy to obtain but $y_i$ more difficult. Assume each $x_i \geq 0$, with not all zero. Suppose that it is desired to estimate the ratio $\sum_i y_i / \sum_i x_i$ without bias and without

[9] or length bias, as it is sometimes called in sampling literature

[10] An unpublished survey by Termeh Shafie on Length-Biased Sampling, found on her ETH webpage, contains a quite useful bibliography.

sampling the entire population. Perhaps $x_i$ is how much the $i^{th}$ customer was billed by their utility company last month, and $y_i$, say a smaller value than $x_i$, the amount they were supposed to have been billed. Suppose we would like to know just how severe the overbilling error is; that is, we would like to know the 'adjustment factor', the ratio $\sum_i y_i / \sum_i x_i$. Even though $\sum_i x_i$ is known, collecting the paired values for everyone is laborious and expensive, so we would like to be able to use a sample of $m < n$ pairs to make an estimate. It is not hard to verify that, if we select a set $R$ of $m$ indices, with all $\binom{n}{m}$ sets equally likely, then the estimate $\sum_{j\in R} y_j / \sum_{j\in R} x_j$ will be biased.

The following device gets around this difficulty. Draw a random set $R$ of size $m$ by first selecting $i$ with size-biased probability $x_i / \sum_j x_j$. Then draw $m-1$ indices uniformly from the remaining $n-1$. Though we are out of the independent framework, the principle of (30) is still at work: size biasing one element has size biased the sum. This is so because we have size biased the one, and then chosen the others from the appropriate conditional distribution. Thus, we have selected a set $r$ of indices with probability proportional to $\sum_{j\in r} x_j$.

From this observation it follows that $\mathbb{E}(\sum_{j\in R} y_j / \sum_{j\in R} x_j) = \sum_j y_j / \sum_j x_j$.

Here is Midzuno's procedure in a bit more detail. Let

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \text{ and } \overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j.$$

First choose index $I$ with distribution

$$P(I=i) = \frac{x_i}{\sum_{j=1}^{n} x_j}.$$

Then from the remaining set $\{1,\ldots,n\}\setminus\{I\}$, take a simple random sample $S$ of size $m-1$. Let $R = S \cup \{I\}$ be the resulting set of size $m$. We claim the estimator $T_R$ is unbiased for $\overline{y}/\overline{x}$, where, for $r$ a subset of $\{1,\ldots,n\}$,

$$T_r = \frac{\overline{y}_r}{\overline{x}_r} \quad \text{with} \quad \overline{y}_r = \frac{1}{m}\sum_{j\in r} y_j \quad \text{and} \quad \overline{x}_r = \frac{1}{m}\sum_{j\in r} x_j.$$

To see why, consider that $R$ may equal $r$, any set of size $m$, in $m$ different possible ways, one each according to first selecting some element $i \in r$ with probability $P(I=i)$, and then collecting the remaining elements in the simple random sample. Hence,

$$\begin{aligned} P(R=r) &= \sum_{i\in r} P(I=i)P(S\setminus\{i\} = r\setminus\{i\}) \\ &= \sum_{i\in r} \frac{x_i}{\sum_{j=1}^{n} x_j} \frac{1}{\binom{n-1}{m-1}} \\ &= \frac{1}{\binom{n-1}{m-1}} \frac{\sum_{i\in r} x_i}{\sum_{j=1}^{n} x_j} \end{aligned}$$

$$= \frac{1}{\frac{n}{m}\binom{n-1}{m-1}} \frac{\overline{x}_r}{\overline{x}}$$
$$= \binom{n}{m}^{-1} \frac{\overline{x}_r}{\overline{x}}.$$

Next, applying the easily shown identity

$$\binom{n}{m}^{-1} \sum_{|r|=m} \overline{y}_r = \overline{y},$$

we obtain

$$\begin{aligned} ET_R &= \sum_{|r|=m} \frac{\overline{y}_r}{\overline{x}_r} P(R=r) = \binom{n}{m}^{-1} \sum_{|r|=m} \frac{\overline{y}_r}{\overline{x}_r} \frac{\overline{x}_r}{\overline{x}} \\ &= \frac{1}{\overline{x}} \binom{n}{m}^{-1} \sum_{|r|=m} \overline{y}_r \\ &= \frac{\overline{y}}{\overline{x}}. \end{aligned}$$

For the variance of the estimator, see [75].

## 5. Relation to Stein's method and concentration inequalities

Implicit in Chen 1975 [25], with improved constants due to [15], see also [45, Theorem 4.12.12], is the following result from [42], Theorem 1.1, see also [77, Theorem 4.10], which we paraphrase[11] here as

**Theorem 5.1.** *Let $X$ be a nonnegative integer valued random variable with $\lambda := \mathbb{E}X \in (0, \infty)$; let $Z$ be Poisson with parameter $\lambda$. Then for any coupling of $X$ with $X^*$, the total variation distance between the distributions of $X$ and $Z$ satisfies*

$$d_{\mathrm{TV}}(X, Z) \leq (1 - e^{-\lambda})\ \mathbb{E}|X^* - (X+1)|.$$

The total variation distance appearing in Theorem 5.1 is defined, for random variables $X, Y$ in general, by $d_{\mathrm{TV}}(X, Y) = \sup_B(\mathbb{P}(X \in B) - \mathbb{P}(Y \in B))$, with the supremum taken over all Borel sets.

Size biasing also has a connection with Stein's method for obtaining error bounds when approximating distributions by the normal distribution, see [14, 13, 26, 43].

[11] The theorem in [42] is stated with the condition that $X$ be a finite sum of indicator random variables. However, an arbitrary nonnegative integer valued $X$ is a sum of indicators, namely $X = \sum_{i \geq 1} 1(X \geq i)$, and the restriction on finite sum can be removed using Theorem 2.3 applied to $X_n := X \wedge n = \sum_{i=1}^{n} 1(X \geq i)$.

Size bias also plays a role in concentration inequalities, see [40, 39, 8, 16]. The results from [40, 8] include: if $X \geq 0$ with $a := \mathbb{E}X \in (0, \infty)$ can be coupled to $X^*$ so that $\mathbb{P}(X^* \leq X + c) = 1$, then

$$\begin{aligned} \text{for } 0 < x \leq a, \mathbb{P}(X \leq x) &\leq (a/x)^{x/c} e^{(x-a)/c} \leq \exp(-(a-x)^2/(2ca)), \\ \text{for } x \geq a, \mathbb{P}(X \geq x) &\leq (a/x)^{x/c} e^{(x-a)/c} \leq \exp(-(x-a)^2/(c(a+x))). \end{aligned}$$

To see how size bias enters, if a coupling satisfies $\mathbb{P}(X^* \leq X + c) = 1$, then for all $x$, the event $X^* \geq x$ is a subset of the event $X \geq x - c$. Hence for $x > 0$,

$$\begin{aligned} x\mathbb{P}(X \geq x) = x\,\mathbb{E}1(X \geq x) &\leq \mathbb{E}(X1(X \geq x)) \\ &= a\,\mathbb{P}(X^* \geq x) \\ &\leq a\,\mathbb{P}(X \geq x - c), \end{aligned}$$

and dividing by $x$ we get

$$\forall x > 0, \quad G(x) \leq \frac{a}{x} G(x - c), \tag{37}$$

Iterating (37) leads to the sharp upper bounds on $\mathbb{P}(X \geq x)$, for each $x \geq a$. An extension to exploit the weaker condition $\mathbb{P}(X^* \leq X + c | X^*) \geq p \in (0, 1)$ is discussed in [32].

In the context of sums of independent random variables each with a bounded range, the concentration bounds based on bounded size bias couplings are stronger than the corresponding Chernoff–Hoeffding bounds, as well as being broader in scope; see [8]. Applications of these bounds to situations involving dependence, such as the number of relatively ordered subsequences of a random permutation, sliding window statistics including the number of $m$-runs in a sequence of coin tosses, the number of local maxima of a random function on a lattice, the number of urns containing exactly one ball in an urn allocation model, and the volume covered by the union of $n$ balls placed uniformly over a volume $n$ subset of $\mathbb{R}^d$, are discussed in [39]. An example showing that the size bias concentration bounds supply a desired uniform integrability, in a situation where the usual Azuma-Hoeffding bounded martingale difference inequality is not adequate, is given in [5].

## 6. Size bias and Palm distributions

The size bias view of arrival times and stationarity, discussed in Section 3, is sometimes expressed in the language of Palm measures for stationary point processes; see [88, Chapter 8] or [34, p. 299] for details. At this level, Palm measures are derived from *simple* point processes, that is, random nonnegative integer valued measures $\xi$ for which any singleton set $\{s\}$ has measure zero or one, and the Palm measure $\xi_s$ corresponds to conditioning on having an arrival at the point $s$.

There is a more general version of Palm measure, which applies to nonnegative random measures; we attribute this to Jagers and Kallenberg, [49, 51, 52]. This version is, quite directly, a generalization of biasing a process $\mathbf{X} = (X_1, X_2, \ldots) \in [0,\infty)^{\mathbb{N}}$ in the direction of its $i$th coordinate, to get $\mathbf{X}^{(i)}$, described in Section 2.3. The setup is: $S$ is a complete separable metric space and $M$ is the set of nonnegative sigma-finite measures on $S$; typical examples include $S = \mathbb{R}$ and $S = \mathbb{R}^d$. Fix a random measure $\xi$, that is, a random element of $M$. The characterizing property of the Palm measures $\xi_s$, for $s \in S$, is that, for bounded measurable functions $g\colon M \to \mathbb{R}$,

$$\mathbb{E}g(\xi_s) = \frac{\mathbb{E}(\xi(ds)g(\xi))}{\mathbb{E}\xi(ds)}. \tag{38}$$

In the restrictive case $S = \mathbb{N}$, a measure $\zeta \in M$ corresponds naturally to the sequence $(z_1, z_2, \ldots) \in [0,\infty)^{\mathbb{N}}$ with $z_i = \zeta(\{i\})$, the mass assigned by the measure to the location $i$ in the underlying space $S$, hence a random measure $\xi$ corresponds to a stochastic process $\mathbf{X} = (X_1, X_2, \ldots)$ with values in $[0,\infty)^{\mathbb{N}}$. In this restrictive case and under this correspondence, with $s = i$, $\xi_s = \mathbf{X}^{(i)}$ is the process $\mathbf{X}$ biased by its $i$th coordinate $X_i$, and $\mathbb{E}\xi(ds) = \mathbb{E}X_i =: a_i$, and (38) looks identical to (20) — the only difference is that in the setup for (20) we needed to *assume* that for each $i$, $\mathbb{E}X_i > 0$ — in particular, one cannot size bias the random variable $X$ which is identically zero. But in the measure context, it would be an unreasonable extra assumption, to require that the intensity measure $\mathbb{E}\xi$ be purely atomic.

The above has fully described a sense in which Palm measures are a generalization of simple size bias. As an application, we provide a solution, in the same spirit, to (part of) Exercise 11.1 in [52]. The exercise asks for a proof of the following theorem, in which the emphasis is that $\xi$ is *not* assumed to be *integer-valued*.

**Theorem 6.1.** *Suppose $\xi$ is a random measure on $S$. For $s \in S$ write $\delta_s$ for deterministic measure "unit mass at $s$". Suppose the Palm measures satisfy: for $s \in S$, $\xi_s = \xi + \delta_s$. Then $\xi$ is a Poisson process.*

**Lemma 6.2.** *Under the hypotheses of Theorem 6.1, for any measurable $B \subset S$ for which $\mathbb{E}\xi(B) \in (0,\infty)$, the random variable $X = \xi(B)$ satisfies $X^* =^d X+1$.*

*Proof.* Write $\mu = \mathbb{E}\xi$ for the intensity measure; this is the deterministic element of $M$ with $\mu(B) = \mathbb{E}(\xi(B))$ for measurable $B \subset S$. With this notation, (38) says

$$\mathbb{E}g(\xi_s) = \frac{\mathbb{E}(\xi(ds)g(\xi))}{\mu(ds)}. \tag{39}$$

The characterization of Palm measures, as written in (39), is shorthand for its multiplied out version,

$$\mathbb{E}g(\xi_s)\ \mu(ds) = \mathbb{E}(\xi(ds)g(\xi)),$$

so that for measurable $B \subset S$,

$$\int_B \mathbb{E}g(\xi_s)\ \mu(ds) = \int_B \mathbb{E}(\xi(ds)g(\xi)) = \mathbb{E}\int_B (\xi(ds)g(\xi)) = \mathbb{E}(g(\xi)\xi(B)).$$

Now fix a measurable $B \subset S$ with $a := \mathbb{E}\xi(B) \in (0,\infty)$, and fix a bounded measurable $f : \mathbb{R} \to \mathbb{R}$. This induces a bounded measurable function $g\colon M \to \mathbb{R}$ via $g(\zeta) = f(\zeta(B))$. This yields $g(\xi) = f(\xi(B))$, so with $X = \xi(B)$ the right side of the display above is $\mathbb{E}(f(X)X)$. From the hypothesis $\xi_s = \xi + \delta_s$ we have, for every $s \in B$, $g(\xi_s) = f(\xi_s(B)) = f((\xi+\delta_s)(B)) = f(\xi(B)+1) = f(X+1)$, and the left side of the display is $\int_B \mathbb{E}g(\xi_s)\ \mu(ds) = \mathbb{E}f(X+1)\int_B \mu(ds) = \mathbb{E}f(X+1)\,\mu(B) = \mathbb{E}f(X+1)\,\mathbb{E}X$. Hence, for bounded measurable $f$, $\mathbb{E}(Xf(X)) = \mathbb{E}f(X+1)\,\mathbb{E}X$. This last relation, now proved for an arbitrary bounded measurable $f : \mathbb{R} \to \mathbb{R}$, shows that $X^* =^d X+1$. □

**Lemma 6.3.** *Suppose that $G$ is an event, and $X$ is Poisson($\lambda$). Write*

$$\begin{aligned} p_i &= \mathbb{P}(X=i), \\ q_i &= \mathbb{E}(1(X=i)1(G)), \\ r_i &= \mathbb{E}(1(X=i)1(G^c)), \end{aligned}$$

*so that $p_i = q_i + r_i$ for $i = 0,1,2,\ldots$. Suppose that for $i = 0,1,2,\ldots$,*

$$(i+1)\,q_{i+1} = \lambda\, q_i, \qquad (i+1)\,r_{i+1} = \lambda\, r_i. \tag{40}$$

*Then $X$ and $G$ are independent.*

*Proof.* The proof is similar to that of Proposition 2.1. In particular, applying induction as there to (40), we obtain $q_i = \mathbb{P}(G)e^{-\lambda}\lambda^i/i! = \mathbb{P}(G)\mathbb{P}(X=i)$. □

*Proof of Theorem 6.1.* Consider a measurable $B \subset S$ for which $\lambda := \mathbb{E}\xi(B) \in (0,\infty)$. Lemma 6.2, combined with Corollary 11.3, shows that $X = \xi(B)$ is Poisson($\lambda$). Now consider an event $G$ which is measurable with respect to the restriction of $\xi$ to $B^c$. Lemma 6.3 shows that $G$ is independent of $X$, with an argument similar to the proof of Lemma 6.2 verifying that the hypotheses of Lemma 6.3 are satisfied. Hence for disjoint subsets $B_1, B_2, \ldots \subset S$ each having $0 < \mathbb{E}\xi(B_i) < \infty$, $\xi(B_1)$ is independent of $(\xi(B_2), \xi(B_3), \ldots)$, so by induction, the Poisson distributed random variables $\xi(B_2), \xi(B_3), \ldots$ are mutually independent. □

**Acknowledgement**

We are grateful to the referee who asked us to consider the the connection between size bias and Palm measures.

## 7. Martingale size bias, and size bias for Galton Watson trees

This section is based mostly on [64], which employs a notion of size-biased Galton Watson trees to give a conceptual and intuitive proof of the Kesten-

Stigum theorem, which we briefly describe below. We also found [79], [35], and [65] useful for clarifying lingering issues involving the "spine" or "backbone" of the size biased Galton Watson trees.

For the reader already familar with the size biased Galton Watson tree, here is a brief description of these issues. a) Is the spine intrinsic to the size-biased tree, or is it just an ingredient in a particular construction? b) Given just the tree, generated using a spine but without labels to show where the spine lies, can the spine be located? c) Can one start with the unbiased Galton Watson tree, and then add a process, of immigrants and their descendants, to get a coupling with a size biased tree? d) If yes to c), can this be done so that the original tree and the difference are independent? We answer a) and b), but leave c) and d) alone.[12]

The Kesten-Stigum theorem concerns the following: Suppose we are given a Galton-Watson branching process with offspring variable $L$, whose distribution is given by $\mathbb{P}(L = k) = p_k$, with mean $m = \sum kp_k \in (0, \infty)$, so that the number of individuals $Z_n$ at time $n$ has $\mathbb{E}Z_n = m^n$. The process given by $W_n := Z_n/m^n$ is a nonnegative martingale, hence converging almost surely to some limit $W$. For $m \leq 1$, it is easy to prove that $Z_n \to 0$ a.s., so $W = 0$; equivalently $\mathbb{E}W = 0$. In particular the martingale is not uniformly integrable. But things are more subtle when $1 < m < \infty$. For this case the Kesten-Stigum theorem asserts that if $\mathbb{E}L \log L < \infty$, then $\mathbb{E}W = 1$, while if $\mathbb{E}L \log L = \infty$, then $\mathbb{E}W = 0$.

The proof of the Kesten-Stigum theorem in [64] begins with the observation that $W_n$ serves as Radon-Nikodym derivative with respect to the usual distribution of $[T]_n$, the branching process tree $T$ observed up to time $n$, of the distribution size-biased by $Z_n$; and since the resulting size biased distributions are consistent, there results a notion of size-biased tree, which we call $T^*$. Namely, this tree is obtained by picking one "special" individual from each generation, and changing its offspring distribution from that of $L$ to that of the size-biased version $L^*$, which satisfies, in particular, $\frac{1}{m}\mathbb{E} \log L^* = \mathbb{E}L \log L$. The Kesten-Stigum criterion is thus whether $\mathbb{E} \log L^*$ is finite or infinite. Write $Y_n$ for the number of *extra* children injected into generation $n+1$ by size-biasing the number of children of the selected special individual from generation $n$; these individuals counted by $Y_n$ (but not their descendants), are called *immigrants*.

It turns out that if $\mathbb{E} \log L^* < \infty$ then the process of immigrants grows subexponentially, so the contribution from immigrants *and* their descendants, up to time $n$, is $O(m^n)$. Then under the size-biased distribution, $W < \infty$ a.s. and the size-biased law of $t$ is absolutely continuous with respect to the original law. On the other hand, if $\mathbb{E} \log L^* = \infty$ then the contribution from immigrants, even without counting their descendants, grows faster than any exponential;

[12] Our reason for leaving c) and d) alone is that there are conflicting notions of the use of immigrants in constructing the size biased tree (though leading, in the end, to the same distribution on trees). In [64], the notion is implicitly established by declaring that the size biased process, with spine removed, is a branching process with immigration — starting with zero individuals, so that every individual is either an immigrant, or else descended from an immigrant. Our construction (48) uses a different notion, leading to a coupling in which there is the original unbiased tree, plus immigrants, plus individuals descended from immigrants.

in particular $W = \infty$ a.s. under the size-biased distribution, and $W = 0$ a.s. under the original distribution. Thus size-biasing plays a natural role in the understanding of this result. See [64] for a proof and further information.

### *7.1. Martingale size bias*

Recall that in Section 2.3 we discussed size-biasing a process $\mathbf{X} = (X_1, X_2, \ldots) \in [0,\infty)^\infty$ with joint law $\mu$, by size-biasing one of its coordinates $X_i$, assuming that $a_i := \mathbb{E}X_i \in (0,\infty)$. The recipe is given by (19), and we wrote $\mathbf{X}^{(i)} = (X_1^{(i)}, X_2^{(i)}, \ldots)$ for the resulting process. A natural question: what is the result if $\mathbf{X}$ is a martingale?

So assume now that $\mathbf{X}$ is a martingale, nonnegative and nonconstant. This implies, in particular, that for each $i = 1, 2, \ldots$, the mean $a_i := \mathbb{E}X_i$ is in $(0,\infty)$, with $a_1 = a_2 = \ldots$; call the common value $a$. For any $i, n \geq 1$, the specifications (19) of the distributions of $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(n)}$, restricted to the first $n$ coordinates, with Radon-Nikodym derivatives expressed in terms of arbitrary bounded measurable $g_n : [0,\infty)^n \to \mathbb{R}$, are that

$$\mathbb{E}g_n(X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)}) = \frac{1}{a}\,\mathbb{E}(X_i\, g_n(X_1, X_2, \ldots, X_n)), \tag{41}$$

and

$$\mathbb{E}g_n(X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)}) = \frac{1}{a}\,\mathbb{E}(X_n\, g_n(X_1, X_2, \ldots, X_n)). \tag{42}$$

By the martingale property of $\mathbf{X}$, for all $i \geq n \geq 1$, the righthand sides of (41) and (42) are equal to each other; hence

$$\text{for } i \geq n \geq 1,\ (X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)}) =^d (X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)}). \tag{43}$$

Now (43) says that we have a consistent family of finite dimensional distributions for a process, which we naturally call a *size-biased martingale*, and which we denote by $\mathbf{X}^* = (X_1^*, X_2^*, \ldots)$. The justification for this notation is that each individual coordinate $X_i^*$ of the process $\mathbf{X}^*$ is a size biased version of $X_i$, in the sense of the original definition (9). The proof, in turn, of this latter statement is that from (43) and the discussion in Section 2.3 we must have $X_n^{(n)} =^d X_n^*$, while (43) applies for every $n$. To recapitulate, the joint distribution of the first $n$ coordinates of $\mathbf{X}^*$ is given by

$$\mathbb{E}g_n(X_1^*, X_2^*, \ldots, X_n^*) = \frac{1}{a}\,\mathbb{E}(X_n\, g_n(X_1, X_2, \ldots, X_n)) \tag{44}$$

for all bounded measurable $g_n : [0,\infty)^n \to \mathbb{R}$. Considering the $n$-th coordinate marginally, the distribution of $X_n^*$ agrees with the elementary definition (9) applied to $X_n$ in the role of $X$. The martingale property of $\mathbf{X}$ plays an essential role in this construction; had $X_1, X_2, \ldots$ been arbitrary non-negative random variables, each with strictly positive finite mean, while the size biased distributions for the $X_1^*, X_2^*, \ldots$ considered individually would still be given by (9), the *joint* distribution is not specified by (9).

Sometimes, one starts with a nonnegative process $\mathbf{Z} = (Z_1, Z_2, \ldots)$ with means $a_i := \mathbb{E}Z_i \in (0, \infty)$, in which the sequence $a_1, a_2, \ldots$, is not constant; but after *scaling out the means* by defining $X_i := Z_i/a_i$, the new process $\mathbf{X} = (X_1, X_2, \ldots)$ turns out to be a martingale. (An example of this is given by $Z_n :=$ the size of the population at time $n$, in any Galton Watson process where the mean number of offspring per individual is $m \in (0, 1) \cup (1, \infty)$.) Then $\mathbf{X}$ can be size biased as above, yielding $\mathbf{X}^* = (X_1^*, X_2^*, \ldots)$. In light of (17), if we set $Z_i^* := a_i X_i^*$ for each $i$, the distribution of this $Z_i^*$ necessarily agrees with the elementary definition (9) of size bias applied to $Z_i$ in the role of $X$, but in addition the joint distribution of the *size biased process* $\mathbf{X}^*$ induces a joint distribution on $\mathbf{Z}^* := (Z_1^*, Z_2^*, \ldots)$, i.e we have obtained a natural *coupling* of the marginal size biased distributions. To recapitulate: given nonnegative random variables $Z_1, Z_2, \ldots$ with $\mathbb{E}Z_i \in (0, \infty)$, and a joint distribution for a process $\mathbf{Z} = (Z_1, Z_2, \ldots)$, if $\mathbf{Z}$ is a martingale, or the process derived from $\mathbf{Z}$ by scaling out the mean motion is a martingale, then there is a process $\mathbf{Z}^*$, simultaneously size biasing every coordinate.

Note that if we start with a martingale $\mathbf{X}$, there is no particular reason for $\mathbf{X}^*$ to be a martingale. Similarly, the *process with mean motion scaled out*, say $\mathbf{Y}$ with $Y_n := X_n^*/\mathbb{E}X_n^*$, need not be a martingale. However, there is an important class for which the martingale-based process bias preserves structure: namely, if the process is *also* a Markov chain, then Markov structure is preserved. We will prove this, in Lemma 7.1.

We limit ourselves to the case where the state space is $\mathbb{Z}_+$, the nonnegative integers, for the sake of easy notation, and also we limit ourselves to the time homogeneous case; neither of these restrictions is essential. To comply with the common convention for indexing time, we switch the index set from $\mathbb{N}$, the natural numbers, to $\mathbb{Z}_+$. And for later application to the special case of Galton Watson processes, we explicitly allow the possibility that state 0 is a trap.

**Lemma 7.1.** *Suppose* $\mathbf{X} = (X_0, X_1, \ldots)$ *is a Markov chain on* $S = \mathbb{Z}_+$ *with transition matrix* $M$*; assume that* $X_0 = 1$*. Suppose that* $r_i := \sum_j j\, M_{ij} < \infty$*, for all* $i \in S$*. Define a new stochastic matrix* $N$*, row by row, by size biasing the* rows *of* $M$*, if possible:*

$$\text{if } r_i > 0, \text{ then } N_{ij} := j\, M_{ij}/r_i; \quad \text{if } r_i = 0, \text{ then } N_{ij} := M_{ij}. \tag{45}$$

*Let* $\mathbf{Y} = (Y_0, Y_1, \ldots)$ *be the Markov chain governed by* $N$*, with* $Y_0 = 1$*. Assume that for all* $n, a_n := \mathbb{E}X_n \in (0, \infty)$*. Let* $W_n := X_n/a_n$*. If* $(W_0, W_1, \ldots)$ *is a martingale, then the size biased process* $\mathbf{X}^*$ *has the same distribution as the Markov chain* $\mathbf{Y}$*.*

Note: the process $\mathbf{Y}$ in the statement of the Lemma may be considered as a special case of *Doob's h-transform*; see [76, p. 296]. The letter $h$ is mnemonic for *harmonic*, and (45) is the special case where $h$ is the identity function — as we pointed out, using the fortuitous choice of the letter $h$ for the identity function, at the start of Section 2.2.

*Proof.* Fix a time $n$ and a sequence $z_0 z_1 \cdots z_n \in S^{n+1}$, with $z_0 = 1$. We need to show that $\mathbb{P}(Y_0 Y_1 \cdots Y_n = z_0 z_1 \cdots z_n) := N_{1z_1} N_{z_1 z_2} \cdots N_{z_{n-1} z_n}$ is equal to $\mathbb{P}(X_0^* X_1^* \cdots X_n^* = z_0 z_1 \cdots z_n) := (z_n/\mathbb{E}X_n)\, \mathbb{P}(X_0 X_1 \cdots X_n = z_0 z_1 \cdots z_n) = (z_n/\mathbb{E}X_n)\, M_{1z_1} M_{z_1 z_2} \cdots M_{z_{n-1} z_n}$. Observe that the martingale hypothesis implies that state 0 is a trap for both processes, i.e., $M_{00} = N_{00} = 1$, and that no other state leads only to 0, i.e., for $i > 0$, $M_{i0} < 1$ and $N_{i0} < 1$, hence $r_i > 0$ and $N_{i0} = 0$. So if some $z_k = 0$, then $z_n = 0$ and using $k$ as the earliest index for which $z_k = 0$, we have $N_{z_{k-1} z_k} = 0$, hence $\mathbb{P}(Y_0 Y_1 \cdots Y_n = z_0 z_1 \cdots z_n) = 0 = \mathbb{P}(X_0^* X_1^* \cdots X_n^* = z_0 z_1 \cdots z_n)$. Otherwise, all $z_i \neq 0$, and the factor for time $k$, of the form $N_{ij}$, is given by $j\, M_{ij}/r_i$, specifically with $i = z_{k-1}, j = z_k$. To use the martingale property for $W$, recall that $X_n = a_n W_n$, and note that $(X_{k-1} = i)$ is the same event as $(W_{k-1} = i/a_{k-1})$. Hence $r_i = \mathbb{E}(X_k | X_{k-1} = i) = \mathbb{E}(a_k W_k | W_{k-1} = i/a_{k-1}) = i\, a_k/a_{k-1}$. Hence the product $N_{1z_1} N_{z_1 z_2} \cdots N_{z_{n-1} z_n}$ telescopes, to the desired value. □

### *7.2. Tree size bias*

Following [64] 'tree' will denote a rooted plane tree, possibly infinite, in which every individual has a finite number, possibly zero, of descendants. We consider the set $\mathcal{T}$ of all trees, and for $t \in \mathcal{T}$ we let $[t]_n$ be the set of all trees whose first $n$ levels agree with $t$. Write $\mathcal{T}_n \subset \mathcal{T}$ for the set of trees of height at most $n$; each $\mathcal{T}_n$ is countable. The sigma algebra $\mathcal{F}_n$ on $\mathcal{T}$ is generated by sets of the form $[t]_n$, $t \in \mathcal{T}_n$, and the sigma algebra $\mathcal{F}$ is generated by the union of the $\mathcal{F}_n$. A probability distribution on $(\mathcal{T}, \mathcal{F})$ can then be specified via a consistent family of probability distributions on $\mathcal{F}_n$, $n = 1, 2, \ldots$.

Write $z_n(t) \geq 0$ for the number of individuals in level $n$ of $t$; for each fixed $n \geq 1$, this gives a nontrivial notion of size. (By our convention that the tree is rooted, we always have $z_0(t) = 1$.) *Any* probability distribution on $\mathcal{T}$, yielding random trees $T$, can be size biased, giving a new distribution yielding trees $T^*$, *provided* that with $Z_n := z_n(T)$, we have both $\mathbb{E}Z_n \in (0, \infty)$ *and* that the process $\mathbf{W} = (W_0, W_1, W_2, \ldots)$ with $W_n := Z_n/\mathbb{E}Z_n$ is a martingale with respect to the filtration $\{\mathcal{F}_n\}$. Specifically, for each $n$ we bias the distribution of trees of height at most $n$, via the following formula: for a given deterministic tree $t_n$ of height at most $n$, with $z_n \geq 0$ individuals at level $n$, we set

$$\mathbb{P}(T^* \in [t_n]_n)) = \frac{z_n}{\mathbb{E}Z_n}\, \mathbb{P}(T \in [t_n]_n)). \tag{46}$$

The proof that the distributions are consistent, and that we have thus defined a tree-valued process, depends on the martingale property of the process $\mathbf{W}$.

### *7.3. The size biased Galton Watson tree, with or without a spine*

Returning to Galton Watson trees, we would like to point out that passing from a Galton-Watson branching *process* to the associated random *tree* depends not

only on the offspring distribution $(p_0, p_1, p_2, \ldots)$, but also an (often implicit) imposition of symmetry. Specifically, let $\mathbf{L} = (L_{n,i})_{n\geq 0, i\geq 1}$ be an array with independent identically distributed entries, where $\mathbb{P}(L_{n,i} = k) = p_k$. The usual recursive construction for the process of *counts*, in which $Z_0 := 1$, and then for $n \geq 0$ we set

$$Z_{n+1} := \sum_{1\leq i\leq Z_n} L_{n,i}, \tag{47}$$

gives rise to a plane tree if we declare that, for $i = 1$ to $Z_n$, the $i$th individual in generation $n$ has $L_{n,i}$ children. The distribution of the plane tree from this standard construction has *maximal* symmetry: at any time $n$, all $Z_n$ subtrees, rooted at an individual of generation $n$, are equal in distribution to the original process.

But alternatively, for certain purposes, given $Z_n = k$ we could sort $L_{n,1}, \ldots, L_{n,k}$ in nonincreasing order, now renamed $A_{n,1} \geq \cdots \geq A_{n,k}$, and then declare that the $i$th individual in generation $n$ has $A_{n,i}$ children. With this construction we would still have the same counts $Z_{n+1}$ as before, and hence the same process $(Z_0, Z_1, Z_2, ...)$, but now a different tree lacking distributional symmetry. Namely, if a parent has more than one child, then his second-born child is guaranteed to produce no more grandchildren than his first-born child produces, and so forth.

By common agreement, the Galton Watson *tree* is the one given by the first construction, with maximal symmetry, rather than the one arising, say, from sorted offspring counts. To size bias this tree let us once again start with (26), which says that a sum of independent (non-negative, finite nonzero mean) random variables is size biased by applying size bias to a single summand. In the spirit of maximal symmetry, we fix one particular joint distribution for $(L, L^*)$, with $L^* \geq L$ always; see (15). Then augment the array $\mathbf{L}$ so that it becomes $\mathbf{L} = ((L_{n,i}, L^*_{n,i}))_{n\geq 0, i\geq 1}$ whose entries are i.i.d. pairs, but possibly with dependence within in each pair. (In Section 11, Theorem 11.2 says precisely when it is possible to have $L$ and $L^* - L$ independent.)

Continuing in the spirit of *maximum* symmetry, to construct the size biased tree $t^*$, one would naturally start with i.i.d. uniform (0,1) random variables $U_0, U_1, \ldots$, independent of $\mathbf{L}$, with $U_n$ used to decide *which* individual in generation $n$ will have a size biased number of children. A little more formally, the tree is constructed recursively: for each $n \geq 0$, given the tree $t^*$ observed up to time $n$, with $Z^*_n$ individuals at time $n$ (and, always $Z^*_n \geq 1$), take $I_n := \lceil Z^*_n U_n \rceil$; then for $i \neq I_n$ the $i^{\text{th}}$ individual in generation $n$ has $L_{n,i}$ children, but for $i = I_n$ the number of children is $L^*_{n,i}$. The resulting tree $t^*$ has

$$Z^*_{n+1} := \sum_{1\leq i\leq Z^*_n} L_{n,i} + (L^*_{n,I_n} - L_{n,I_n}), \tag{48}$$

and by (26), this is equal in distribution to $(\sum_{1\leq i\leq Z^*_n} L_{n,i})^*$ — note that the sum, being size-biased, has $Z^*_n$ i.i.d. summands, rather than $Z_n$ summands as in (47).

**Exercise 7.2.** *Check that the distribution of the tree produced by the above procedure has distribution satisfying* (46).

To be complete, without spoiling the reader's fun, we supply a solution but postpone it until the end of Section 7.5. For historical reasons, we hereby name the tree from above procedure as the *spineless* (biased) tree.

In contrast to the maximal symmetry spineless procedure described above, and following [64], for $n \geq 1$, we could restrict the $n$th generation candidates for size bias, instead of all $Z_n$ individuals, to just the $S_{n-1} := L^*_{n-1,V_{n-1}}$ children of the individual $V_{n-1}$ in generation $n-1$ who was size biased. So in this tree $V_0 = 1$ and, for $n > 0$, $V_n$ is descended from $V_{n-1}$; and the non backtracking path from the root, $(V_0, V_1, V_2, \ldots)$ is called the *spine* of the biased tree. We refer to the tree in this construction as the *spinal* (biased) tree. To recapitulate: the spineless tree has a list $(I_0 = 1, I_1, I_2, \ldots)$ of biased individuals, i.e., person $I_n$ in generation $n$ uses the distribution of $L^*$ to dictate his unusually large number of children. By contrast, while the spinal tree has a similar list of biased individuals, $(V_0 = 1, V_1, V_2, \ldots)$, with $V_n$ in generation $n$, *these* biased nodes form a path. The procedure with a spine can be traced back to [57] Kesten 1986, who was studying critical GW processes, conditional on nonextinction, and used the term *backbone* instead of *spine*.

Here are two natural questions:

**Question** 1 Given a random spineless tree and a random spinal tree, without being told which individuals are biased, can one tell which tree came from which procedure?

**Question** 2 Given a random spinal tree, without being told where the spine is, can one identify the spine?

Question 2 is easily answered for subcritical or critical Galton Watson proceses, since in these cases there is a unique infinite path. We will say more about these cases in Section 7.6.

Initially, we found it hard to guess the answer to Question 1; it is *not obvious* whether the spineless tree and the spinal tree have the *same* distribution. But since a computation confirms that the spinal tree also satisfies (46), the answer to Question 1 is a definite *no*: while the two procedures have different joint distributions for (tree, bias markers), they have the same marginal distribution for tree.

**Exercise 7.3.** *Derive the marginal distribution for spinal trees. Hint, it may help to show, at the same time, that conditional on the tree up to time $n$, with $k = Z^*_n$ individuals at level $n$, the spinal position $V_n$ in generation $n$ is uniformly distributed from* 1 *to* $k$.

The marginal distribution for spinal trees is derived in [64], and we give a derivation, at the end of Section 7.5.

The answer to Question 2, for supercritical processes, depends, as does the Kesten-Stigum result, on whether $\mathbb{E}L \log L$ is finite or infinite. We spend the rest of this section on this dichotomy, and then return, in section 7.6, to consideration of the subcritical and critical cases.

**Theorem 7.4.** *Consider a supercritical Galton Watson process, with offspring distribution $L$ having $\mathbb{E}L < \infty$, and the size biased tree generated by the spinal procedure. Given the tree alone:*

1. *If $\mathbb{E}L \log L = \infty$, the spine can be correctly identified, with probability 1.*
2. *If $\mathbb{E}L \log L < \infty$, any procedure to find the spine fails, with probability 1.*

Before giving the proof, we remark that case (1) is easy, because $\mathbb{E}L \log L = \infty$ implies that the distributions of tree and size-biased tree are mutually singular. In the other case, with $\mathbb{E}L \log L < \infty$, the distribution of biased tree is absolutely continuous with respect to the unbiased tree, but some work is needed. We need something similar to Fano's inequality, giving a lower bound on the error probability for classification, but Fano requires the Kullback-Liebler divergence to be finite. So we are led to prove two lemmas about selection in a general setting.

### *7.4. Selecting one special item out of k choices, assuming $Q \ll P$*

We want to detect the one item sampled from $Q$, when mixed in with $k-1$ others sampled from $P$.

Lemmas 7.5 and 7.6 below are stated and proved for a fairly general pair of distributions $P$ and $Q$ satisfying $Q \ll P$, meaning that $Q$ is dominated by $P$, i.e., $Q$ is absolutely continuous with respect to $P$, i.e., $P(A) = 0$ implies $Q(A) = 0$. We will apply Lemma 7.6 to a Galton Watson tree for which $\mathbb{E}L \log L$ is finite and $m > 1$; then $P$ will be the GW tree law, and $Q$ the size-biased law. We remark that though case 1) in the proof of Lemma 7.6 cannot occur in the Galton Watson situation, nonetheless we prefer to have Lemma 7.6 in its natural generality.

**Setup for selecting the special one, out of $k$ choices.** Fix $k > 1$. Let $P, Q$ be laws on a Polish space $S$. Let $Y_1, \ldots, Y_k$ be independent, with $Y_1$ sampled from $Q$, and $Y_2, \ldots, Y_k$ sampled from $P$, and let $X_1, \ldots, X_k$ be obtained from the $Y$s by an independent uniformly distributed random permutation $\pi \in \mathcal{S}_k$. So $X_i = Y_{\pi(i)}$, and then $I := \pi(1)$ identifies the index of the $X$ value sampled from $Q$; one might write $\mathbf{X} = \mathbf{Y} \circ \pi$. A *selection procedure* is a function $f : S^k \to [k]$, meant as a guess of $I$ as a function of the sample $\mathbf{X} = (X_1, \ldots, X_k)$. The *score* for a selection procedure $f$ is $s(f) := \mathbb{P}(f(\mathbf{X}) = I)$.

In the case $Q \ll P$, it is "obvious" that the best selection procedures, i.e., those with maximal score, are precisely those which inspect the likelihood ratio $r(x) = dQ/dP(x)$, and pick $I$ arbitrarily from those indices $i$ which maximize $r(X_i)$, relative to the $k$-sample. To prove this, while taking into account possible ties, we define a particular candidate $f_0$ for best selection procedure, by picking the *earliest* index among those $i$ for which $r(X_i) = \max(r(X_1), \ldots, r(X_k))$.

**Lemma 7.5. Optimal selection**. *In the setup above,* any *selection procedure $f : S^k \to [k]$ satisifes*

$$s(f) \leq s(f_0),$$

*and furthermore $s(f) = s(f_0)$ implies that with $J := f(\mathbf{X})$ we have $r(X_J) = \max(r(X_1), \ldots, r(X_k))$ with probability one.*

*Proof.* Write $\mathbf{x} = (x_1, \ldots, x_k)$ and $z = r(x_1) + \cdots + r(x_k)$. Conditional on $\mathbf{X} = \mathbf{x}$, the odds $[\mathbb{P}(I = 1) : \mathbb{P}(I = 2) : \cdots : \mathbb{P}(I = k)]$ are equal to $[r(x_1) : r(x_2) : \cdots : r(x_k)]$, hence $\mathbb{P}(I = i|\mathbf{X} = \mathbf{x}) = r(x_i)/z$. Thus $\mathbb{P}(f_0(\mathbf{X}) = I|\mathbf{X} = \mathbf{x}) = \max(r(x_1), \ldots, r(x_k))/z$; any competing procedure $f$ has $\mathbb{P}(f(\mathbf{X}) = I|\mathbf{X} = \mathbf{x}) = r(x_{f(\mathbf{x})})/z$, with the same denominator $z$, and no larger a numerator, hence $\mathbb{P}(f(\mathbf{X}) = I|\mathbf{X} = \mathbf{x}) \leq \mathbb{P}(f_0(\mathbf{X}) = I|\mathbf{X} = \mathbf{x})$ with equality holding if $r(x_{f(\mathbf{x})}) = r(x_{f_0(\mathbf{x})})$. Taking expectation yields the inequality $s(f) \leq s(f_0)$, and the claim regarding when $s(f) = s(f_0)$. □

**Lemma 7.6. Lost in the noise**. *Let $P$ and $Q$ be distinct probability distributions on $S$, with $Q$ dominated by $P$. Given $\varepsilon > 0$, there exists $k_0 < \infty$ such that for all $k \geq k_0$, in the setup above, with $Y_1$ distributed according to $Q$ and $Y_2, \ldots, Y_k$ distributed according to $P$, every selection procedure $f$ has $s(f) < \varepsilon$.*

*Proof.* Using Lemma 7.5, we may assume that the selection procedure is $f_0$, choosing an item of maximal likelihood ratio $r$.

Take any version $r$ of the Radon-Nikodym derivative, $r(x) = dQ/dP(x)$; our hypothesis implies that $\mathbb{E}r(Y_2) = 1$, and $r$ is $Q$-almost surely finite, i.e., $\mathbb{P}(r(Y_1) < \infty) = 1$. Let $u \in (0, \infty]$ be the essential sup of $r$; we get the same essential sup with respect to $P$ and with respect to $Q$. We deal with two separate cases. Informally, in case 1, the special item might achieve the maximal value, i.e., $r(Y_1) = u$, but even so, there is likely to be a many-way tie against noise. In case 2, the value $u$ is unobtainable, and most likely, some nonspecial choice strictly beats the special, i.e., there is some random $J \in [2, k]$, with $r(Y_J) > r(Y_1)$. More formally:

Case 1: $p := \mathbb{P}(r(Y_1) = u) > 0$. Note, this implies that $u < \infty$, hence $\mathbb{P}(r(Y_2) = u) = p/u > 0$. So we pick $k_0$ so that, if $N$ is distributed Binomial$(k_0 - 1, p/u)$, then $\mathbb{P}(N \leq 2/\varepsilon) < \varepsilon/2$. Hence the event that no more than $2/\varepsilon$ items in the sample have $r(X_i) = u$ contributes at most $\varepsilon/2$ to $\mathbb{P}(f(\mathbf{X}) \neq I)$, and on the complementary event, by exchangeability, the conditional probability of picking $I$ correctly is less than $\varepsilon/2$.

Case 2: $\mathbb{P}(r(Y_1) = u) = 0$. We can pick $t < u$ so that $q := \mathbb{P}(r(Y_1) \geq t) \in (0, \varepsilon/2)$. Note that $\mathbb{P}(r(Y_2) \geq t) \geq q/u > 0$. When $r(Y_1) < t$ and $k$ is large, with high probability at least one of the $k - 1$ items generated from $P$ will have a higher value for $r$ than that of the item generated from Q. That is, by taking $k_0$ large enough, so that $(1 - q/u)^{k_0 - 1} < \varepsilon/2$, we can guarantee that for all $k \geq k_0$, $\mathbb{P}($ at least one of $Y_2, \ldots, Y_k$ has $r(Y_i) \geq t) > 1 - \varepsilon/2$. Hence a procedure, like $f_0$, that picks one item from among those achieving $\max(r(X_1), \ldots, r(X_k))$, has probability at least $1 - \varepsilon$ of picking incorrectly. □

### *7.5. Proof of the spinal identification dichotomy*

**Proof of Theorem 7.4**. Identification of the spine $(V_0, V_1, V_2, \ldots)$ is the conjunction of identifying $V_n$, for $n \geq 0$. So with respect to the 0–1 dichotomy,

for (1) it suffices to show that for each $n$, $V_n$ can be correctly identified with probability 1, while for (2), it suffices to show that for arbitrary $\varepsilon > 0$, there exists $n = n(\varepsilon)$ for which the probability of correct identification of $V_n$ is less than $\varepsilon$.

For the spinal tree, observed up to time $n$, and conditional on the event $Z_n^* = k$, the distribution of the $k$ rooted subtrees with roots at time $n$ fits exactly the setup described in Section 7.4: one of the trees is distributed according to $Q$, the law of the size biased GW tree, the other $k-1$ are distributed according to $P$, the unbiased GW tree law, all $k$ are mutually independent, and, using the uniformity of $V_n$ in $[k]$, as proved in [64], the joint distribution of these $k$ trees matches that of the random permutation $\pi$ applied to an ordered sample $Y_1, \ldots, Y_k$ in which $Y_1$ has the distribution $Q$.

For (1), with $\mathbb{E}L \log L = \infty$, Theorem A in [64] includes the statement that $P$ and $Q$ are mutually singular. So pick a subset $A \subset \mathcal{T}$ having $P(A) = 0, Q(A) = 1$, and given the $k$ rooted subtrees, pick the node whose subtended tree lies in $A$, thereby finding $V_n$ correctly with probability 1.

For (2), with $\mathbb{E}L \log L < \infty$, Theorem A in [64] includes the statement that $Q \ll P$. Given $\varepsilon > 0$, apply Lemma 7.6 to find a value $k_0$ that works for $\varepsilon/2$. Then use the supercriticality to find a single value $n$ for which $\mathbb{P}(Z_n \geq k_0 | Z_n > 0) > 1 - \varepsilon/2$. The combination shows that, given all the subtrees rooted at time $n$, the chance of correctly picking the one whose root is $V_n$ is less than $\varepsilon$. Then, since conditional on the tree up to time $n$, the location of $V_n$ was uniform from 1 to $Z_n^*$, and using conditional independence of past and future, given the present, any function applied to the entire tree, attempting to identify $V_n$, has probability less than $\varepsilon$ of gettng the correct value.
QED

**Answers to the exercises on spineless and spinal trees**. Suppose the given tree $t_n$ of height at most $n$ has $z_j$ nodes at height $j$, for $j = 0$ to $n$. Write $GW(t_n)$ for the Galton Watson probability that the tree observed up to time $n$ matches this tree, corresponding to the last factor on the right side of (46). For both the spineless tree and the spinal tree, we calculate the joint distribution of tree and bias markers up to time $n$, then sum over possible locations of the bias markers, to show that the marginal distribution of tree satisfies (46).

**Answer to Exercise 7.2** For the spineless procedure, recall our notation $I_n$ designating which individual in generation $n$ gets biased, and write $\mathbf{I} = (I_0, I_1, I_2, \ldots, I_{n-1})$ for the process naming those individuals arising in forming a tree up to time $n$. The possible values for $\mathbf{I}$ form a set $S$, with $|S| = z_0 z_1 \cdots z_{n-1}$, and using the notation $[k] := \{1, 2, \ldots, k\}$, $S$ is the Cartesian product $S = [z_0] \times [z_1] \times \cdots \times [z_{n-1}]$. Given both $t_n$ and $\mathbf{i} \in S$, so that we specify the tree up to time $n$, *and* which nodes were biased, at stages 0 to $n-1$, write $k_0, k_1, \ldots, k_{n-1}$ for the respective offspring counts for the biased nodes. For the joint probability of tree and bias markers, taking into account first size bias factors of the form $\mathbb{P}(L^* = k)/\mathbb{P}(L = k) = k/m$, and then factors of the

form $1/z_j = \mathbb{P}(I_j = i_j)$, and using $z_0 = 1$, we have

$$\mathbb{P}(T^* \in [t_n]_n), \mathbf{I} = \mathbf{i}) = \frac{k_0}{m} \frac{k_1}{m} \cdots \frac{k_{n-1}}{m} \; GW(t_n) \; \frac{1}{z_1 z_2 \cdots z_{n-1}}.$$

When we sum over $S$ to get the marginal distribution of tree up to time $n$, the sum factors as a product indexed by time $j = 0$ to $n-1$, and the $k_j$ values sum to $z_{j+1}$, explicitly $k_{j,1} + k_{j,2} + \cdots + k_{j,z_j} = z_{j+1}$, yielding

$$\mathbb{P}(T^* \in [t_n]_n) = \sum_{\mathbf{i} \in S} \mathbb{P}(T^* \in [t_n]_n), \mathbf{I} = \mathbf{i}) = \frac{z_1}{m} \frac{z_2}{m} \cdots \frac{z_n}{m} \; GW(t_n) \; \frac{1}{z_1 z_2 \cdots z_{n-1}}.$$

$$= \frac{z_n}{m^n} \; GW(t_n),$$

which is (46) in the Galton-Watson case.

**Answer to Exercise 7.3** For the spinal procedure, write $V_0, V_1, \ldots, V_n$ for the random spine, for the tree restricted up to time $n$. Given the tree, the initial segments of spine, down through times $0, \ldots, n$, are in one-to-one correspondence with the nodes $V_0, V_1, \ldots, V_n$ along the spine. Given both $t_n$, and $v_n \in [z_n]$ to serve as the value of $V_n$, so that we specify the tree up to time $n$, *and* which nodes were biased, yielding a path $v_0, v_1, \ldots, v_n$ from root to $v_n$, write $k_0, k_1, \ldots, k_{n-1}$ for the respective offspring counts for the biased nodes. For the joint probability of tree and bias markers, taking into account first size bias factors of the form $\mathbb{P}(L^* = k)/\mathbb{P}(L = k) = k/m$, and next factors of the form $1/k_j = \mathbb{P}(V_{j+1} = v_{j+1})$, we have

$$\mathbb{P}(T^* \in [t_n]_n), V_n = v_n)) = \frac{k_0}{m} \frac{k_1}{m} \cdots \frac{k_{n-1}}{m} \; GW(t_n) \; \frac{1}{k_0 k_1 \cdots k_{n-1}}$$
$$= \frac{1}{m^n} \; GW(t_n).$$

Summing over the $z_n$ possible values for $v_n$, to give the marginal distribution of tree up to time $n$, shows that the spinal tree satisfies (46).

### *7.6. Subcritical and critical GW, conditional on survival forever*

For a Galton Watson process, consider the event of survival forever, that is, $A := \{\forall n, Z_n > 0\}$. If the process is subcritical — $0 < m < 1$, or critical — $m = 1$, then $\mathbb{P}(A) = 0$. Conditioning on $A$, by definition, means taking the limit, as $n \to \infty$, of conditioning on $Z_n > 0$. Athreya-Ney [12, pp. 58] prove that, when $\mathbb{P}(A) = 0$, conditioning on $A$ achieves the same distribution as size biasing, although their result imposes the extra hypothesis that $p_1 := \mathbb{P}(L = 1) > 0$. In this section, we give an elementary proof, without the extra hypothesis.

**Lemma 7.7.** *Suppose that $q$ and $q(\ell)$ are probability measures on $\mathbb{N}$, (indexed by $\ell \in \mathbb{N}$ or $\ell \in \mathbb{R}$) so that, in particular, $q_j \geq 0$ for all $j \in \mathbb{N}$, and $1 = \sum_{j \geq 1} q_j$. Let $S := \{j : q_j > 0\}$ be the support of $q$, and assume that $S$ is also the support*

*of $q(\ell)$, for every $\ell$. Suppose that as $\ell \to \infty$, the $q(\ell)$-odds converge to the $q$-odds, that is*

$$\forall j, k \in S, \ \frac{q_j(\ell)}{q_k(\ell)} \to \frac{q_j}{q_k}. \tag{49}$$

*Suppose also that*

$$\text{the family } \{q(\ell)\} \text{ is tight.} \tag{50}$$

*Then $q(\ell) \to^d q$, equivalently,*

$$\forall j \in \mathbb{N}, \ q_j(\ell) \to q_j. \tag{51}$$

*Proof.* By tightness, every subsequence of the $q(\ell)$ has a subsubsequence with a limit. Pick such a subsubsequence, and call its limit $p$. Along this subsubsequence, $q_j(\ell) \to p_j$ and $q_k(\ell) \to p_k$; if $p_k > 0$ then $p_j/p_k = \lim q_j(\ell)/q_k(\ell) = q_j/q_k$, for every $j \in S$. This implies $p = q$, and of course all convergent subsequential limits being the same $q$ implies that $q(\ell) \to^d q$ as $\ell \to \infty$. □

**Lemma 7.8.** *Take the same setup as Lemma 7.7, assuming (49), but in place of (50), supposing instead that*

$$\forall j, k \in S \text{ with } j > k, \ \frac{q_j(\ell)}{q_k(\ell)} \nearrow \frac{q_j}{q_k}, \tag{52}$$

*where the upward arrow denotes convergence upward. Then (50) holds, so the conclusion (51) holds.*

*Proof.* Using (52), for any $k \in S$

$$\frac{\sum_{j>k} q_j(\ell)}{\sum_{j\le k} q_j(\ell)} \le \frac{\sum_{j>k} q_j}{\sum_{j\le k} q_j}.$$

Given $\varepsilon > 0$, pick $k \in S$ so that the right side above is less than $\varepsilon$. This implies that for every $\ell$, the left side is also less than $\varepsilon$, which implies the tightness hypothesis (50). □

**Theorem 7.9.** *Let $\mathbf{Z}$ be a subcritical or critical Galton Watson process, so that the offspring distribution $L$ has $m := \mathbb{E}L \in (0, 1]$. Then the size biased process, $\mathbf{Z}^*$, is equal in distribution to $\mathbf{Z}$ conditional on survival forever; equivalently, as $n \to \infty$, $(\mathbf{Z}|Z_n > 0) \to^d \mathbf{Z}^*$.*

*Proof.* The core of the proof is the asymptotic relation, for fixed $k$, $(1 - (1 - \delta)^k) \sim k\,\delta$ as $\delta \to 0$.

Fix a time $n > 0$ and a value $i > 0$ with $\mathbb{P}(Z_{n-1} = i) > 0$. We use Lemma 7.1 with the Galton Watson process serving as the Markov process whose transition matrix is $M$; row $i$ of $M$ gives the distribution of $Z_n$ conditional on $Z_{n-1} = i$, and size biasing leads to row $i$ of $N$ giving the distribution of $Z_n^*$ conditional on $Z_{n-1}^* = i$. For use in Lemma 7.8, we take $q$ to be the distribution on $\mathbb{N}$ given by row $i$ of $N$, as specified by (45), and we take $q(\ell)$ to be the distribution of $Z_n$ conditional on ($Z_{n-1} = i$ and $Z_{n+\ell} > 0$). Writing $\delta(\ell) := \mathbb{P}(Z_\ell > 0)$,

the probability of survival for an additional $\ell$ units of time, starting from a population of size 1, we have, with *proportional to* denoted by $\propto$,

$$\begin{aligned} q_k(\ell) &:= \mathbb{P}(Z_n = k | Z_{n-1} = i, Z_{n+\ell} > 0) \\ &= M_{ik} \, \frac{1 - (1 - \delta(\ell))^k}{1 - (1 - \delta(\ell+1))^i} \\ &\propto M_{ik} \, (1 - (1 - \delta(\ell))^k). \end{aligned} \tag{53}$$

Using the hypothesis that the GW process is subcritical or critical, $\delta(\ell) \searrow 0$ as $\ell \to \infty$. This easily implies the hypothesis (52) for Lemma 7.8. So $(\mathbf{Z}|A)$ is a Markov process, whose transition matrix is $N$, and this process, called $\mathbf{Y}$ in Lemma 7.1, is equal in distribution to $\mathbf{Z}^*$, using both the martingale and Markov properties of GW, but not the full structure of GW, to enable Lemma 7.1. □

Next consider a subcritcal or critical Galton Watson tree, conditional on survival forever. Thanks to (46), combined with Theorem 7.9, it is "obvious" that the conditioned tree is the size biased tree. A proof can be found in [1], and we now present a more direct proof.

**Theorem 7.10.** *Consider the tree $T$ for a subcritical or critical Galton Watson process, so that the offspring distribution $L$ has $m := \mathbb{E}L \in (0, 1]$. The size biased tree $T^*$, as specified by (46), is equal in distribution to $T$ conditional on survival forever; equivalently, as $n \to \infty$, $(T | Z_n > 0) \to^d T^*$.*

*Proof.* Fix $i$, consider the associated $q$ and $q(\ell)$ from the proof of Theorem 7.9, and recall the notation $\delta(\ell)$ for $\mathbb{P}(Z_\ell > 0)$, the probability of survival for an additional $\ell$ units of time, starting from a population of size 1. By formula (51) and Lemma 7.8 we know that $q_k(\ell) \to q_k$ as $\ell \to \infty$, for $k$ in the support of $q$. Therefore, using (53), we see that

$$M_{i,k} \, \frac{k\delta(\ell)}{i\delta(\ell+1)} \to N_{i,j} = M_{i,k} \, \frac{k}{i\, m}$$

Thus we must have $\delta(\ell)/\delta(\ell+1) \to 1/m$ as $\ell \to \infty$, and hence

$$\text{for fixed } n, \ \delta(\ell)/\delta(\ell+n) \to 1/m^n \text{ as } \ell \to \infty. \tag{54}$$

Now for fixed $n > 0$ and $t \in \mathcal{T}$, with $k := z_n(t)$, using (54),

$$\begin{aligned} \mathbb{P}(T \in [t]_n | Z_{n+\ell} > 0) &= \mathbb{P}(T \in [t]_n) \, \frac{1 - (1 - \delta(\ell))^k}{\delta(n+\ell)} \\ &\to \mathbb{P}(T \in [t]_n) \, \frac{k}{m^n}, \end{aligned}$$

in the limit as $\ell \to \infty$. Comparison with (46) completes the proof. □

## 8. Size bias, tightness, and uniform integrability

Recall that a collection of random variables $\{Y_\alpha : \alpha \in I\}$, where $I$ is an arbitrary index set, is *tight* iff for all $\varepsilon > 0$ there exists $L < \infty$ such that

$$\mathbb{P}(Y_\alpha \notin [-L, L]) < \varepsilon \quad \text{for all } \alpha \in I.$$

This definition looks quite similar to the definition of uniform integrability, where we say $\{X_\alpha : \alpha \in I\}$ is *uniformly integrable*, or UI, iff for all $\delta > 0$ there exists $L < \infty$ such that

$$\mathbb{E}(|X_\alpha|; X_\alpha \notin [-L, L]) < \delta \quad \text{for all } \alpha \in I.$$

Intuitively, tightness for a family is that uniformly over the family, the probability mass due to large values is arbitrarily small. Similarly, uniform integrability is the condition that, uniformly over the family, the contribution to the expectation due to large values is arbitrarily small. Since *size bias* relates contribution to the expectation to probability mass, it should be possible to use size bias to express a relation between uniform integrability and tightness.

We show, in Theorems 8.1 and 8.2, that for random variables, i.e., real valued random *elements*, there is an intimate connection between tightness and uniform integrability, and that this connection is made via size bias. But we must note, the concept of tightness is much broader than the concept of uniform integrability, in that tightness applies to random elements of metric and topological spaces, whereas uniform integrability is inherently a real valued notion. In more general spaces, to define tightness, the closed intervals $[-L, L]$ are replaced by arbitrary *compact sets*, and the discussion below relates only to metric spaces with the property that balls $\{x : d(x, y) \le L\}$ are compact.

To discuss the connection between size biasing and uniform integrability, it is useful to restate the basic definitions in terms of nonnegative random variables. It is clear from the definition of tightness above that a family of *nonnegative* random variables $\{Y_\alpha : \alpha \in I\}$ is tight iff for all $\varepsilon > 0$ there exists $L < \infty$ such that

$$\mathbb{P}(Y_\alpha > L) < \varepsilon \quad \text{for all } \alpha \in I, \tag{55}$$

and from the definition of UI, that a family of *nonnegative* random variables $\{X_\alpha : \alpha \in I\}$ is uniformly integrable iff for all $\delta > 0$ there exists $L < \infty$ such that

$$\mathbb{E}(X_\alpha; X_\alpha > L) < \delta \quad \text{for all } \alpha \in I. \tag{56}$$

For general random variables, the family $\{G_\alpha : \alpha \in I\}$ is tight [respectively UI] iff $\{|G_\alpha| : \alpha \in I\}$ is tight [respectively UI]. Hence we specialize in the remainder of this section to random variables that are non-negative.

Care must be taken to distinguish between the *additive* contribution to expectation, and the *relative* contribution to expectation. The following example makes this distinction clear. Let

$$\mathbb{P}(X_n = n) = 1/n^2, \mathbb{P}(X_n = 0) = 1 - 1/n^2, \quad n = 1, 2, \ldots.$$

Here, $\mathbb{E}X_n = 1/n$, the family $\{X_n\}$ is uniformly integrable, but $1 = \mathbb{P}(X_n^* = n)$, so the family $\{X_n^*\}$ is not tight; the additive contribution to the expectation from large values of $X_n$ is small, but the *relative* contribution is large — one hundred percent! The following two theorems, which exclude this phenomenon, show that tightness and uniform integrability are very closely related.

**Theorem 8.1.** *Assume that for $\alpha \in I$, where $I$ is an arbitrary index set, the random variables $X_\alpha$ satisfy $X_\alpha \geq 0$ and $0 < \mathbb{E}X_\alpha < \infty$, and let $Y_\alpha =^d X_\alpha^*$. Then*

$$\{X_\alpha : \alpha \in I\} \text{ is UI if } \{Y_\alpha : \alpha \in I\} \text{ is tight.}$$

*Assume further that the values $\mathbb{E}X_\alpha$ are uniformly bounded away from 0, say $c > 0$ and $\forall \alpha, c \leq \mathbb{E}X_\alpha$. Then*

$$\{X_\alpha : \alpha \in I\} \text{ is UI iff } \{Y_\alpha : \alpha \in I\} \text{ is tight.}$$

*Proof.* Since $Y_\alpha =^d X_\alpha^*$, by (9), for every $L$ we have $\mathbb{P}(Y_\alpha > L) = \mathbb{E}(1(Y_\alpha > L)) = \mathbb{E}(X_\alpha 1(X_\alpha > L))/\mathbb{E}X_\alpha$, so

$$\mathbb{E}(X_\alpha; X_\alpha > L) = \mathbb{E}X_\alpha \, \mathbb{P}(Y_\alpha > L).$$

First, we show that tightness implies UI. Assume that $\{Y_\alpha : \alpha \in I\}$ is tight, and take $L_0 > 0$ to satisfy (55) with $\varepsilon = 1/2$, so that $\mathbb{P}(Y_\alpha > L_0) < 1/2$ for all $\alpha \in I$. Hence, for all $\alpha \in I$,

$$\mathbb{E}(X_\alpha; X_\alpha > L_0) = \mathbb{E}X_\alpha \mathbb{P}(Y_\alpha > L_0) < \mathbb{E}X_\alpha/2,$$

and therefore,

$$\begin{aligned} L_0 \geq \mathbb{E}(X_\alpha; X_\alpha \leq L_0) &= \mathbb{E}X_\alpha - \mathbb{E}(X_\alpha; X_\alpha > L_0) \\ &> \mathbb{E}X_\alpha - \mathbb{E}X_\alpha/2 = \mathbb{E}X_\alpha/2, \end{aligned}$$

and hence $\mathbb{E}X_\alpha < 2L_0$. Now given $\delta > 0$ let $L$ satisfy (55) for $\varepsilon = \delta/(2L_0)$. Hence $\forall \alpha \in I$,

$$\mathbb{E}(X_\alpha; X_\alpha > L) = \mathbb{E}X_\alpha \, \mathbb{P}(Y_\alpha > L) < 2L_0 \, \mathbb{P}(Y_\alpha > L) < 2L_0 \, \varepsilon = \delta,$$

establishing (56).

Second we show that UI implies tightness, in the presence of means bounded uniformly away from zero . Assume that $\{X_\alpha : \alpha \in I\}$ is UI, and let $\varepsilon > 0$ be given to test tightness in (55). Let $L$ be such that (56) is satisfied with $\delta = \varepsilon c$. Now, using $\mathbb{E}X_\alpha \geq c$, for every $\alpha \in I$,

$$\mathbb{P}(Y_\alpha > L) = \mathbb{E}(X_\alpha; X_\alpha > L)/\mathbb{E}X_\alpha \leq \mathbb{E}(X_\alpha; X_\alpha > L)/c < \delta/c = \varepsilon,$$

establishing (55). □

As an alternate to Theorem 8.1, for the sake of having cleaner hypotheses and a cleaner conclusion, we also give the following theorem. Note below that the $X_\alpha$ to be involved in size bias are allowed to have $\mathbb{E}X_\alpha = 0$ — it is not a typo — because we will be taking $(X_\alpha + c)^*$ for some $c > 0$.

**Theorem 8.2.** *Assume that for $\alpha \in I$, where $I$ is an arbitrary index set, the random variables $X_\alpha$ satisfy $X_\alpha \geq 0$ and $\mathbb{E}X_\alpha < \infty$. Pick any $c \in (0, \infty)$, and for each $\alpha$ let $Y_\alpha = (c + X_\alpha)^*$. Then*

$$\{X_\alpha : \alpha \in I\} \text{ is UI iff } \{Y_\alpha : \alpha \in I\} \text{ is tight.}$$

*Proof.* By Theorem 8.1, the family $\{c + X_\alpha\}$ is UI iff the family $\{(c + X_\alpha)^*\}$ is tight. As it is easy to verify that the family $\{X_\alpha\}$ is tight [respectively UI] iff the family $\{c + X_\alpha\}$ is tight [respectively UI], Theorem 8.2 follows directly from Theorem 8.1. □

## 9. Size bias, the lognormal, and Chihara–Leipnik

In this section we review a construction due to Chihara in 1970, [27], and Leipnik in 1979, [59, 60], of a family of discrete distributions having the same moment sequence as the lognormal. Durrett [37] presents this result with the comment "Somewhat remarkably, there is a family of discrete random variables with these moments." We hope here to show that, from the point of view of size bias, this construction is *natural* and *inevitable*, but we can only speculate that for the original discoverers, size bias played a role in the creative process, perhaps via (10); see [60, page 332, formula (16)]. As a reward for using size bias, we are able to show, in Theorem 9.4, that the lognormal itself is a mixture of these discrete distributions, and furthermore that these discrete distributions are the extreme points of a Choquet simplex — in this case, the set $U_c$ of solutions of (65), which is a subset of the closed convex set $V_c$ formed by all distributions having the same moments as the lognormal $X = \exp(\sqrt{\log c}\, Z)$. The results in this section, linking the lognormal distribution with size bias, appear in [29, 69, 68]; see also [62].

Throughout this section, we write $Z$ for a standard normal, with moment generating function $M(\beta) = e^{\beta^2/2}$. The *standard lognormal* is given by $X = e^Z$, with moments

$$\mathbb{E}X^n = \mathbb{E}\exp(n\,Z) = M(n) = e^{n^2/2}, \tag{57}$$

for $n = 0, 1, 2, \ldots$. (It is clear that (57) holds for all $n \in (-\infty, \infty)$, but historically, moments usually refer to the case $n = 0, 1, 2, \ldots$.) Similarly, for $\sigma > 0$, the lognormal $X = e^{\sigma Z}$ obtained by exponentiating the normal with mean zero and variance $\sigma^2$ has moments $\mathbb{E}X^n = \mathbb{E}\exp(n\,\sigma Z) = M(\sigma n) = e^{n^2\sigma^2/2}$. Hence it is natural to define, taking $c = e^{\sigma^2} \in (1, \infty)$,

$$V_c := \{\mu : \mu = \mathcal{L}(X) \text{ for some } X \geq 0, \text{ with } \mathbb{E}X^n = c^{n^2/2}, n \geq 0\}. \tag{58}$$

The famous fact that $V_c$ is not a singleton set, i.e., that the lognormal distribution is not determined by its moments, is from Stieltjes in 1894 [83, Section 56, page J. 106], reprinted in [84]. The family of examples in (61) is also from Stieltjes [83], although probabilists, e.g., [37, 38], attribute it to Heyde, who rediscovered it in 1963 [48]. These alternate probability distributions having the

same moments as the lognormal are continuous, with density presented via a perturbation of the lognormal density, as follows. We will write $f_{0,\sigma^2}$ for the density of the lognormal $X = e^{\sigma Z}$:

$$f_{0,\sigma^2}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp(-(\log x)^2/(2\sigma^2)), \ x \in (0,\infty). \tag{59}$$

For positive integers $m$ and real $\delta \in [-1,1]$ define

$$g_{m,\delta}(x) = 1 + \delta \sin(2\pi m \log x/\sigma^2), \ x \in (0,\infty), \tag{60}$$

so in case $\delta = 0$, one has $g_{m,\delta}(x) = 1$ for all $x$. Let $h_{m,\delta}$ be given by

$$h_{m,\delta}(x) = f_{0,\sigma^2}(x) \times g_{m,\delta}(x), \ x \in (0,\infty). \tag{61}$$

One then checks that for integers $n$, $\int x^n h_{m,\delta}(x)\ dx = \int x^n f_{0,\sigma^2}(x)\ dx = e^{n^2\sigma^2/2}$. In particular, the case $n = 0$ shows that $h_{n,\delta}$, clearly a non-negative function, is a density.

Let $X = e^Z$, and consider its size biased version, $X^*$. By (14) and (57), for integers $n$,

$$\mathbb{E}(X^*)^n = \frac{\mathbb{E}X^{n+1}}{\mathbb{E}X} = \frac{M(n+1)}{M(1)} = e^n M(n) = e^n \mathbb{E}X^n = \mathbb{E}(eX)^n. \tag{62}$$

Of course, since the lognormal distribution is *not* characterized by its moments, this only *suggests*, and does *not prove*, that $X^* =^d eX$. Similarly, the general lognormal and its moments are given by

$$X = \exp(\sigma Z + \mu), \ \mathbb{E}X^n = e^{\mu n + \sigma^2 n^2/2} \tag{63}$$

and calculation of the moments of $X^*$ *suggests* that for $X = \exp(\sigma Z + \mu)$ we have $X^* =^d e^{\sigma^2} X$. Simple computation with the density and (4) shows that indeed,

$$X = \exp(\sigma Z + \mu) \text{ has } X^* =^d cX, \text{ with } c = \exp(\sigma^2). \tag{64}$$

We leave the proof of (64) as an exercise for the reader, with our solution given by this[13] footnote.

As regards the distributional family, varying $\mu$ corresponds to scaling $X$, and $X \mapsto yX$ is a trivial transformation, so it makes sense to study only the case $\mu = 0$. But varying $\sigma$ is nontrivial; it corresponds to taking ordinary powers, $X \mapsto (X)^\sigma$. So, we fix $\mu = 0$ and let $\sigma > 0$ be arbitrary. We will write $c = \exp(\sigma^2) > 1$; alongside our standard notation, $a = \mathbb{E}X$, for $X = e^{\sigma Z}$ we have $a = \mathbb{E}X = \sqrt{c}$.

[13]The density of $\exp(\mu + \sigma Z)$ is $f_{\mu,\sigma^2}(x) = 1/(x\sqrt{2\pi}\sigma) \exp(-(\log x - \mu)^2/(2\sigma^2))$. Expanding the square in the exponent, and keeping track only of factors that vary with $x$, we have $f_{\mu,\sigma^2}(x) \propto (1/x)x^{-(\log x)/(2\sigma^2)} x^{\mu/\sigma^2}$. Hence for any real $\beta$, $f_{\mu+\beta\sigma^2,\sigma^2}(x) \propto x^\beta f_{\mu,\sigma^2}(x)$. The case $\beta = 1$ shows that $x f_{\mu,\sigma^2}(x)$ is proportional to $f_{\mu+\sigma^2,\sigma^2}(x)$, hence by (4), $(\exp(\mu + \sigma Z))^* =^d \exp((\mu + \sigma^2) + \sigma Z)$.

For the remainder of this section, for $c \in (1, \infty)$, we investigate random variables satisfying

$$X \geq 0, \qquad \mathbb{E}X = \sqrt{c}, \qquad X^* =^d cX, \tag{65}$$

along with the corresponding set of probability distributions,

$$U_c := \{\mu : \mu = \mathcal{L}(X), \text{ for some random variable X, satisfying (65)}\}. \tag{66}$$

With this notation, (63) and (64) assert that $X = \exp(\sqrt{\log c}\, Z)$ satisfies (65), and its lognormal distribution is an element of $U_c$.

As the first step in our investigation of (65), inspired by Feynman's maxim,[14] we note that our considerations lead twice to a homogenous system of equations, of the form

$$\forall n \in \mathbf{Z}, s_{n+1} = ac^n s_n, \quad \text{which has solution} \quad s_n = s_0 a^n c^{n(n-1)/2}. \tag{67}$$

For the first instance of (67), write $m_n := \mathbb{E}X^n$, with $m_1 = \mathbb{E}X = a$, so the moment shift relation (14) can be written as $\mathbb{E}(X^*)^n = m_{n+1}/a$. Using (65), we have $\mathbb{E}(X^*)^n = \mathbb{E}(cX)^n = c^n\, m_n$, hence

$$m_{n+1} = ac^n\, m_n. \tag{68}$$

Combining $m_0 = 1$ with the solution to (67), we have

$$m_n = a^n c^{n(n-1)/2} = c^{n^2/2} \text{ (using } a = \sqrt{c}), \tag{69}$$

for all $n \in \mathbf{Z}$. In summary, so far we have shown that $U_c \subset V_c$, i.e., any solution of (65) has the same moments as the lognormal $e^{\sigma Z}$.

For the second instance of (67), if $X$ satisfying (65) has any pointmass at some $b > 0$, then it must have pointmass at every point $b\, c^n$ for $n \in \mathbf{Z}$. With the benefit of hindsight[15] we go doubly negative, and for $n \in \mathbf{Z}$ define $p_n$ and $r_n$ by $r_n = 1/p_n = 1/\mathbb{P}(X = bc^{-n})$. We have $p_{n+1} = \mathbb{P}(X = bc^{-n-1}) = \mathbb{P}(cX = bc^{-n}) = \mathbb{P}(X^* = bc^{-n}) = (bc^{-n}/a)\mathbb{P}(X = bc^{-n}) = (bc^{-n}/a)p_n$, so that

$$r_{n+1} = (a/b)c^n\, r_n. \tag{70}$$

This is (67) with $r_n$ in the role of $s_n$ and $a/b$ in the role of $a$, so quoting the solution, and using $a = \sqrt{c}$, we get $p_0/p_n = r_n/r_0 = (a/b)^n c^{n(n-1)/2} = b^{-n} c^{n^2/2}$. Finally, replacing $n$ by $-n$ in $p_n = \mathbb{P}(X = bc^{-n})$, we have, for $n \in \mathbf{Z}$,

$$\mathbb{P}(X = bc^n) = b^{-n} c^{-n^2/2}\ \mathbb{P}(X = b). \tag{71}$$

With some fixed $c > 1$ in mind, for any $b \in (0, \infty)$ we call the set

$$\{\ldots, b/c^2, b/c, b, bc, bc^2, bc^3, \ldots\}$$

[14] Feynman Lectures on Physics, Vol. 2, Chapter 12.1 and oft again, "the same equations have the same solutions."

[15] Defining for example $s_n = \mathbb{P}(X = bc^n)$ or $s_n = 1/\mathbb{P}(X = bc^n)$ or $s_n = \mathbb{P}(X = bc^{-n})$ does not lead directly to (67) — the reader *might* enjoy trying these.

the "orbit of $b$," for short, or to say it fully, the orbit of $b$ modulo multiplication by powers of $c$. The language here comes from the theory of a group acting on a set; orbits are equivalence classes, and $(0,\infty)$ is a disjoint union of orbits. For a set containing exactly one representative for each orbit, the natural choice is $[1,c)$.

If we want $X$ supported on a single orbit, that is, with $1 = \sum_{n\in\mathbf{Z}} p_n$, then we need

$$\mathbb{P}(X = bc^n) = b^{-n}c^{-n^2/2}/t(b,c), \text{ where } t(b,c) := \sum_{m\in\mathbf{Z}} b^{-m}c^{-m^2/2}. \tag{72}$$

The function $t$ is essentially the Jacobi theta function; the convergence of the series, for any $c > 1$, is obvious.

However, the calculation connecting (71) with (65) was done *assuming* that $\mathbb{E}X = \sqrt{c}$, and we will only have succeeded, in getting a random variable with $X^* =^d cX$ and supported on a single orbit, if, and only if, it turns out that, under the mass function (72), one has $\mathbb{E}X = \sqrt{c}$. (It is trivial to check that if (71) *and* $\mathbb{P}(X \in \{\ldots, b/c^2, b/c, b, bc, bc^2, bc^3, \ldots\}) = 1$ *and* $\mathbb{E}X = \sqrt{c}$, then (65) is true.) So crossing our fingers we calculate, from (72),

$$\mathbb{E}X/\sqrt{c} = \sum_{n\in\mathbf{Z}} bc^n \; b^{-n}c^{-n^2/2} \; c^{-1/2}/\, t(b,c) = 1,$$

with the change of variables $m = n - 1$ justifying the final equality.

The above discussion shows how the use of size bias, particularly (65), makes it relatively straightforward to rediscover and prove the following theorem of Chihara and Leipnik:

**Theorem 9.1** (Chihara – Leipnik)**.** *For any $\sigma > 0$, with $c := \exp(\sigma^2)$, and for any $b \in (0,\infty)$, there is a distribution $\ell(b,c)$ for a discrete random variable $X_{b,c}$, whose support is the single orbit $\{\ldots, b/c^2, b/c, b, bc, bc^2, bc^3, \ldots\}$, with probability mass function given by (72). This random variable satisfies (65), which implies that for $n \in \mathbf{Z}$, $\mathbb{E}X_{b,c}^n = \exp(n^2\sigma^2/2)$, so taking $n \geq 0$ in particular, the discrete random variable $X_{b,c}$ has the same moments as the lognormal $\exp(\sigma Z)$, where $Z$ is standard normal.*

Another issue is whether the lognormal can be expressed as a mixture of these discrete distributions. Leipnik 1991, [60, page 337], wrote[16] "*One hopes that for some mixing distribution $dh(b)$ we have* that the lognormal distribution for $e^{\sigma Z}$ is a mixture, governed by $h$, of the single orbit distributions, *and so too*

$$\phi(t) = \int_0^\infty \phi_b(t) \; dh(b).$$

[The display above expresses the characteristic function of the lognormal as a mixture of the characteristic functions of the distributions $\ell(b,c)$.] *Unfortunately, the necessary $d\,h(b)$ is somewhat complicated and hence sheds little light*

[16] Italics to show Leipnik's exact words, and ordinary text to show our paraphrase.

*on the sum distribution problem. However, the extraordinary non-uniqueness of the lognormal moment problem is apparent.*

The words "one hopes" signal a conjecture; the sentence beginning "Unfortunately $\cdots$" suggests that he may have had a proof too messy to publish. Whatever the case, we supply a proof here, in the form of Theorem 9.4 below. Conceivably, the complication encountered by Leipnik might have arisen from considering mixtures indexed by $(0, \infty)$, without exploiting the formula $\ell(b, c) = \ell(bc, c)$ — the proof of which we leave as an exercise for the reader. It is natural, and simple, to take mixtures indexed by $[1, c)$; then there is a unique choice for $h$, with one simple computation to check. For notation, we follow Leipnik, and write $dh(b)$ to denote a general measure $h$ to govern a mixture; so that $h$ may be discrete, absolutely continuous, singular continuous, or a mixture of these. In the special case in Theorem 9.4 given by (76), expressing the lognormal as a mixture of the $\ell(b, c)$, we have $h$ absolutely continuous, with density $h_c$ with respect to Lebesgue measure.

We show how to express the lognormal as a mixture of the Chihara–Leipnik discrete distributions $\ell(b, c)$ from Theorem 9.1, via Lemma 9.2, Lemma 9.3, and Theorem 9.4. There is a related result, expressing a particular continuously distributed random variable, not the lognormal, but having the same moments, as a mixture of these discrete distibutions, in [17, Proposition 2.2].

**Lemma 9.2.** *Fix $c > 1$. For any probability measure $h$ on $[1, c)$, the mixture of the laws $\ell(b, c)$, governed by $dh(b)$, gives a distribution for $X$ which satisfies (65). The set $U_c$ of distributions which satisfies (65) is closed and convex, hence* any *mixture of distributions which satisfy (65) is also a distribution which satisfies (65).*

*Proof.* Since for each $b \in [1, c)$, $m(b) := \mathbb{E}X_{b,c} = \sqrt{c}$, we are in the situation for Lemma 2.4 where the measure $h'$ governing $X^*$ as a mixture of the $X^*_{b,c}$ is the same as the original $h$, governing $X$ as a mixture of the $X_{b,c}$. Hence (65) holds, since, obviously, scaling respects mixtures, i.e., the law of $cX$ is the mixture, governed by $h$, of the laws of $c\,X_{b,c}$.

That $U_c$ is *closed* is a bit subtle. Assume we are given $X_1, X_2, \ldots$ with each $X_n$ satisfying (65), and that $X_n \Rightarrow Y$. Obviously $cX_n \Rightarrow cY$, and Theorem 2.3 asserts that $X_n^* \Rightarrow Y^*$, which combined with (17) and (65) gives $Y^* =^d cY$. But (65) also demands that $\mathbb{E}Y = \sqrt{c}$, so one must know that the family $\{X_1, X_2, \ldots\}$ is *uniformly integrable*. Fortunately, (69) implies that $\mathbb{E}X^2 = c^2$ for any solution of (65), which implies that the family is uniformly integrable.

Finally, for the convexity of $U_c$, just as with mixtures of the $\ell(b, c)$, Lemma 2.4 applies, with the same measure governing $X$ as mixture of solutions $X_\alpha$, governing $X^*$ as a mixture of the $X^*_\alpha$, and $cX$ as a mixture of the $cX_\alpha$. □

**Lemma 9.3.** *Suppose $c > 1$, and $X, Y$ are positive random variables which satisfy*

$$0 < \mathbb{E}X = \mathbb{E}Y < \infty \text{ and } X^* =^d cX, Y^* =^d cY.$$

*If the laws of $X$ and $Y$, both restricted to $[1, c)$ agree, even only up to a constant*

*mass factor $k \geq 0$, i.e., if*

$$\text{for all measurable } A \subset [1,c),\ \mathbb{P}(X \in A) = k\,\mathbb{P}(Y \in A). \tag{73}$$

*then $X =^d Y$. (The case $k = 0$ is specifically included in the hypothesis (73), but in every case, the conclusion implies that $k = 1$.)*

*Proof.* Let $a := \mathbb{E}X$, so by hypothesis, we also have $a = \mathbb{E}Y$, (but unlike (65), we are *not* assuming that $a = \sqrt{c}$). Let $S(n)$ be the statement that for all bounded measurable $g$ which vanish outside $[c^n, c^{n+1})$, we have $\mathbb{E}g(X) = k\,\mathbb{E}g(Y)$. The hypothesis (73) clearly implies the statement $S(0)$. Assume now that $S(n)$ holds. Given a bounded measurable function $g$ which vanishes off of $[c^{n+1}, c^{n+2})$, we *define* new functions $g', g''$ by $g'(x) = g(x)/x$ and $g''(x) = g'(cx)$. Clearly $g''$ is bounded, and vanishes off of $[c^n, c^{n+1})$. We have

$$\mathbb{E}g(X) = \mathbb{E}(Xg'(X)) = a\,\mathbb{E}g'(X^*) = a\,\mathbb{E}g'(cX) = a\,\mathbb{E}g''(X)$$

and similarly $\mathbb{E}g(Y) = a\,\mathbb{E}g''(Y)$. Invoking $S(n)$ for the function $g''$, we get

$$\mathbb{E}g(X) = a\,\mathbb{E}g''(X) = ak\,\mathbb{E}g''(Y) = k\mathbb{E}g(Y), \tag{74}$$

hence $S(n)$ implies $S(n+1)$.

A similar argument shows that $S(n)$ implies $S(n-1)$. In detail, given a bounded measurable function $g$ which vanishes off of $[c^{n-1}, c^n)$, we *define* new functions $g', g''$ by $g'(x) = g(x/c)$, so that $g'(cx) = g(x)$, and $g''(x) = xg'(x)$. Clearly $g''$ is bounded, and vanishes off of $[c^n, c^{n+1})$. We have

$$\mathbb{E}g(X) = \mathbb{E}(g'(cX)) = \mathbb{E}g'(X^*) = \frac{1}{a}\mathbb{E}(Xg'(X)) = \frac{1}{a}\mathbb{E}g''(X)$$

and similarly $\mathbb{E}g(Y) = (1/a)\mathbb{E}g''(Y)$; hence (74) holds exactly as before, but this time showing that $S(n)$ implies $S(n-1)$.

Finally, knowing $S(n)$ for all $n \in \mathbf{Z}$ implies that for bounded measurable $g$, $\mathbb{E}g(X) = k\,\mathbb{E}g(Y)$, and the special case $g = 1$ shows that $k = 1$, and hence $X =^d Y$. □

The following Theorem 9.4 applies in particular to the case where $X$ has the lognormal distribution with density $f(x) = 1/(x\sqrt{2\pi}\sigma)\exp(-(\log x)^2/(2\sigma^2))$, recalling that with $c = \exp(\sigma^2)$, $X$ satisfies (65).

**Theorem 9.4.** *Let $X$ be any positive random variable which satisfies (65). Then there is a unique probability measure $h$ on $[1,c)$ such that the distribution of $X$ is the mixture, governed by $dh(b)$, of the Chihara–Leipnik single orbit distributions $\ell(b,c)$ of Theorem 9.1, with point mass functions (and Jacobi theta function $t$) given by (72). The measure $h$ governing the mixture is specified as follows: let $B$ be distributed as $X$, conditional on $(X \in [1,c))$. Then the probability measure $h$ has Radon-Nikodym derivative, relative to the distribution of $B$, given by*

$$\frac{h(db)}{\mathbb{P}(B \in db)} = \frac{t(b,c)}{\mathbb{E}t(B,c)}. \tag{75}$$

*If the distribution of $X$ is absolutely continuous with respect to Lebesgue measure, so that $X$ has a density $f$, the recipe (75) says that with $t$ given by (72), and normalizing constant $k_c$ and function $h_c$ with domain $[1, c)$, defined by*

$$k_c := \int_{x=1}^{c} f(x)\, t(x, c)\ dx, \quad h_c(b) := \frac{1}{k_c} f(b)\, t(b, c), \tag{76}$$

*the measure $h$ governing the mixture has density $h_c$, so that for measurable $A \subset [1, c)$, $h(A) = \int_{b \in A} h_c(b)\ db$.*

*Proof.* First, we must show that the distribution of $B$ was well-defined, i.e., that $\mathbb{P}(X \in [1, c)) > 0$. Here we argue by contradiction: if $\mathbb{P}(X \in [1, c)) = 0$, then Lemma 9.3 could be invoked, with $Y = e^{\sigma Z}, k = 0$, to prove $X =^d Y$, a contradiction since $\mathbb{P}(Y \in [1, c)) > 0$.

Now write $Y$ for a random variable whose distribution is the mixture of the $\ell(b, c)$, governed by $h$. We use the Dirac notation, that $\delta_x$ is unit mass at $x$, so that $\int g(z)\delta_x(dz) = g(z)$ for any measurable $g$. Restricting our attention to $b \in [1, c)$, the Chihara–Leipnik distributions are then expressed as

$$\ell(b, c) = \sum_{n \in \mathbf{Z}} \mu_{b,n} \quad \text{where } \mu_{b,n} := \frac{b^{-n} c^{-n^2/2}}{t(b, c)}\ \delta_{bc^n}.$$

so that $\mu_{b,n}$ is the measure $\ell(b, c)$ restricted to the interval $[c^n, c^{n+1})$ — this uses $b \in [1, c)$.

Focus on the case $n = 0$, so that $\mu_{b,0}$ is mass $1/t(b, c)$ at the point $b$. The specification of $h$ in (75) implies directly that the hypothesis (73) holds — with $k = \mathbb{P}(X \in [1, c)) \times \mathbb{E}t(B, c)$. Hence by Lemma 9.3, we have $X =^d Y$.

The argument for uniqueness is essentially the same: suppose that $Y$ is a mixture of $\ell(b, c)$, governed by some probability measure $h$ on $[1, c)$, and that $X =^d Y$, not assuming that $h$ is given by (75). Restricting the distributions of both $X$ and $Y$ to $[1, c)$, it is clear, from $\mu_{b,0} = 1/t(b, c)\ \delta_b$, that the Radon-Nikodym derivative $h(db)/\mathbb{P}(X \in db\,|X \in [1, c))$ must be proportional to $t(b, c)$. The recipe in (75) gives the unique constant of proportionality to make such an $h$ into a probability measure. □

For each $c > 1$, Lemma 9.2 says that the convex set spanned by the Chihara–Leipnik distributions $\ell(b, c)$, $b \in [1, c)$ is a *subset* of the set $U_c$ of all solutions of (65). Theorem 9.4 asserts that $U_c$ is spanned by the $\ell(b, c)$, so together with the obvious property that any single $\ell(b, c)$ is not a nontrivial mixture of other $\ell(b', c)$, one now knows that the *extreme* points of the set of solutions of (65) *are* the distributions $\ell(b, c)$, for $b \in [1, c)$. For historical naming and perspective: Choquet's Theorem states that for a convex compact subset of a normed space, every point can be represented as a mixture, governed by a probability measure, of extreme points; this probability measure need not be unique, even in the finite dimensional setting. However, in the finite dimensional setting, uniqueness holds when the convex set is a simplex. In honor of this, a convex set, for which the every point has a unique representation as a mixture of extreme points, is

called a *Choquet simplex*. The additional information in Theorem 9.4 about the *uniqueness* of $h$ is then summarized by saying that the set $U_c$ of solutions to (65) forms a *Choquet simplex*.

It is now natural to ask whether Stietljes' examples, with density given by (61), lie in this Choquet simplex.

**Proposition 9.5.** *For every $\sigma > 0$, integer $m$, and real $\delta \in [-1, 1]$, the random variable $X$ with density given by Stieltjes' formula (61) satisfies $X^* =^d cX$, with $c = \exp(\sigma^2)$, and hence $X$ satisfies (65).*

*Proof.* For random variables with a density, the size bias scaling relation in (65), can be expressed in terms of the density, as follows. First, when $X$ has density $f$, the scaled multiple $cX$ has density $(1/c)f(x/c)$. Second, when $X$ has density $f$, and mean $a = \mathbb{E}X = \sqrt{c}$, (4) states that $X^*$ has density $(x/\sqrt{c})f(x)$. Hence, if $X$ has density $f$, mean $\sqrt{c}$, and

$$\forall x \in (0, \infty), \;\; f(x/c) = x\sqrt{c} f(x), \tag{77}$$

then $X$ satisfies (65). Now it is clear that (77) holds for $f = h_{m,\delta}$ given by (61): we have $c = e^{\sigma^2}$, and upon substituting $x/c$ for $x$, the lognormal factor $f_{0,\sigma^2}$ supplies the factor $x\sqrt{c}$, and the perturbation factor $g_{m,\delta}$ supplies no change, since dividing $x$ by $c$ causes $\log x$ to decrease by $\log c = \sigma^2$, so that the argument to the sine function, $2\pi m \log x/\sigma^2$, goes down by $2\pi m$. □

To review: both the lognormal *and* the examples given by Stieltjes are solutions of (65) and hence lie in the Choquet simplex $U_c$. Do all distributions having the lognormal moment sequence lie in this simplex, i.e., does $U_c = V_c$? Berg [18, Proposition 2.1], proved $U_c \subsetneq V_c$ by exhibiting elements of $V_c \setminus U_c$. These distributions can be described as the perturbations of the Chihara–Leipnik distribution in (72) by a factor of $(1 + s(-1)^n)$, for $s \in [-1, 1]$. In detail, Berg showed that for any $c > 1$,

$$b = \sqrt{c}, s \in \{-1, 1\}, \;\; \mathbb{P}(X_s = bc^n) = (1+s(-1)^n)b^{-n}c^{-n^2/2}/t(b, c), n \in \mathbf{Z} \tag{78}$$

leads to $\mathbb{E}X_s^n = c^{n^2/2}$ for $n \in \mathbf{Z}$. In particular, for $b = \sqrt{c}$ the Chihara–Leipnik distribution $\ell(b, c)$ is the midpoint of the line connecting the distributions of $X_{-1}$ and $X_1$. The construction is special to $b = \sqrt{c}$ as the only value of $b \in [1, c)$ for which a line of distributions with moments $\mathbb{E}X^n = c^{n^2/2}$ can be constructed, with $\ell(b, c)$ as the midpoint.

We have shown that $U_c$ is a Choquet simplex; the question as to whether $V_c$ is a Choquet simplex is open. We thank Christian Berg, private communication, for this information and several references, and also for correcting two erroneous conjectures from an earlier draft of our paper.

## 10. Size bias and Skorohod embedding

Skorohod's embedding theorem states that given a nonconstant mean zero random variable $X$, there is a random time $T$ for Brownian motion $(W_t)_{t \geq 0}$ such

that $X =^d W_T$. We discuss Skorohod's proof as presented, for example, in [37, 67]. The proof is based on the construction of a joint distribution for a dependent pair $(U, V)$ with $U, V \geq 0$ so that, with the pair independent of the Brownian motion, the random time $T := T_{U,V} := \inf\{t : W_t \notin [-U, V]\}$ yields $X =^d W_T$. Since $\mathbb{P}(W_T = V|U, V) = U/(U + V))$, and the function $(u, v) \mapsto u/(u + v)$ is nonlinear, it is somewhat surprising that a simple distribution of $(U, V)$ can satisfy $X =^d W_{T_{U,V}}$. That distribution, specified in [37, 67] by the formula

$$dH_\mu(u, v) = (v - u)1(u \leq 0 \leq v)\ \mu(du)\mu(dv)/\mathbb{E}X^+,$$

where $\mu$ is the distribution of $X$, is the same[17] as the distribution (80) below in our size bias treatment. Display (82) highlights how size bias overcomes the nonlinearity of $(u, v) \mapsto u/(u+v)$. The excellent survey by Obłój [67] should be consulted for the history and connections to *the potential of a measure.*

To define the joint distribution for $(U, V)$ in $[0, \infty)^2$, consider random variables $A, B$ with values in $[0, \infty)$ with distribution given by

$$\mathcal{L}(A) = \mathcal{L}(-X|X < 0), \quad \mathcal{L}(B) = \mathcal{L}(X|X > 0);$$

since $X$ is nonconstant and mean zero, both $p_- := \mathbb{P}(X < 0) > 0$ and $p_+ := \mathbb{P}(X > 0) > 0$, so the conditioning is elementary. Note that

$$\mathbb{E}A = \mathbb{E}X^-/p_-, \quad \mathbb{E}B = \mathbb{E}X^+/p_+, \text{ and } \mathbb{E}X^- = \mathbb{E}X^+. \tag{79}$$

Write $p_0 := \mathbb{P}(X = 0)$. Since $A$ and $B$ have finite positive mean, the size biased distributions of $A^*$ and $B^*$ are well defined. Couple so that $A, A^*, B, B^*$ are independent. The final recipe, writing $\delta_q$ for unit mass at the point $q$, is

$$\mathcal{L}(U, V) = p_+\ \mathcal{L}(A^*, B) + p_0\ \delta_{(0,0)} + p_-\ \mathcal{L}(A, B^*), \tag{80}$$

and then take $(U, V)$ to be independent of the Brownian motion $W$.

To prove that (80) and $T = T_{U,V}$ achieve $X =^d W_T$, first consider the case where $\mathbb{P}(X = 0) = 0$. Given a bounded measurable function $h : \mathbb{R} \to \mathbb{R}$, conditioning on $U, V$ and using the exit distribution for Brownian motion from the interval $[-u, v]$ we have

$$\mathbb{E}(h(W_T)|U = u, V = v) = h(-u)\frac{v}{u+v} + h(v)\frac{u}{u+v} =: g(u, v). \tag{81}$$

Next, since we are in the case where $p_- + p_+ = 1$, using (79) we have

$$p_- = \frac{\mathbb{E}B}{\mathbb{E}A + \mathbb{E}B}, \quad p_+ = \frac{\mathbb{E}A}{\mathbb{E}A + \mathbb{E}B}.$$

The size bias relation for processes from Section 2.3, together with the independence of $A, B$, justifies the transition from line 2 to line 3 below: for any

[17] apart from a notational switch between $-u$ and $u$; we write $-u \leq 0 \leq v$ and they write $u \leq 0 \leq v$.

bounded measurable $g : \mathbb{R}^2 \to \mathbb{R}$,

$$\begin{aligned} \mathbb{E}g(U,V) &= p_+ \, \mathbb{E}g(A^*, B) + p_- \, \mathbb{E}g(A, B^*) \\ &= \frac{\mathbb{E}A \, \mathbb{E}g(A^*, B) + \mathbb{E}B \, \mathbb{E}g(A, B^*)}{\mathbb{E}(A+B)} \\ &= \frac{\mathbb{E}(Ag(A,B)) + \mathbb{E}(Bg(A,B))}{\mathbb{E}(A+B)} \\ &= \frac{\mathbb{E}((A+B) \; g(A,B))}{\mathbb{E}(A+B)}. \end{aligned} \tag{82}$$

Using this identity for our function $g$ defined in (81), and using the independence of $A$ and $B$ to go from line 3 to line 4, we have

$$\begin{aligned} \mathbb{E}h(W_T) &= \mathbb{E}g(U,V) \\ &= \frac{\mathbb{E}((A+B) \; g(A,B))}{\mathbb{E}(A+B)} \\ &= \frac{\mathbb{E}\,(h(-A)B + h(B)A)}{\mathbb{E}(A+B)} \\ &= \frac{\mathbb{E}B}{\mathbb{E}(A+B)} \, \mathbb{E}h(-A) + \frac{\mathbb{E}A}{\mathbb{E}(A+B)} \, \mathbb{E}h(B) \\ &= p_- \, \mathbb{E}h(-A) + p_+ \, \mathbb{E}h(B) \\ &= \mathbb{E}(h(X)|X<0) \; p_- + \mathbb{E}(h(X)|X>0) \; p_+ \\ &= \mathbb{E}h(X), \end{aligned}$$

and hence $\mathcal{L}(W_T) = \mathcal{L}(X)$, as claimed.

That $X =^d W_T$ in the general situation, allowing $\mathbb{P}(X=0) \in (0,1)$, is easily seen, since the distribution of $X$ is then a mixture of pointmass at zero, and the distribution of $X$ conditional on $X \neq 0$, and the recipe (80) is the corresponding mixture of pointmass at (0,0) and the distribution of $(U,V)$ treated above.

## 11. Size bias and infinite divisibility

Paul Lévy's theory of infinitely divisible distributions is celebrated; see any of [21, 30, 38, 53] for introductory treatments, or [3, 19, 78] for advanced treatments. For the special case of nonnegative random variables with finite mean, size bias provides an easy handle on the theory.

### *11.1. Steutel revisited*

**Theorem 11.1.** *Suppose $X$ can be size biased, i.e., $X \geq 0$ and $a := \mathbb{E}X \in (0,\infty)$. If $X$ is infinitely divisible, then there exists a distribution for $Y$ such that*

$$X^* =^d X + Y, \quad \text{and } X, Y \text{ are independent.} \tag{83}$$

*Conversely, given that $X$ can be size biased, and that (83) holds for some $Y$, then $X$ is infinitely divisible.*

*In either case, the distribution of $Y$ is unique, and $\mathbb{P}(Y \geq 0) = 1$.*

Remark: In [81] (see also [80]), F. Steutel shows that a cumulative distribution function $F$ on $[0, \infty)$ is infinitely divisible iff it satisfies

$$\int_0^x u dF(u) = \int_0^x F(x-u) dK(u)$$

for a non-decreasing $K$. Our decomposition (83) is clearly a consequence of his integral formula, though he does not use the language of size biasing –he does not, in fact, assume that $F$ has finite mean– and his proof proceeds by way of the Levy representation formula, which we will derive instead as a corollary of (83). Steutel's result is also presented in Sato [78], Theorem 51.1, as well as in the book [82] by Steutel and van Harn.

*Proof.* We begin by assuming that $X$ is infinitely divisible, which by definition means that for each $n$ there exists a distribution such that if $X_1^{(n)}, \ldots, X_n^{(n)}$ are i.i.d. with this distribution, then

$$X =^d X_1^{(n)} + \cdots + X_n^{(n)}. \tag{84}$$

Then by (30)

$$X^* =^d (X - X_1^{(n)}) + (X_1^{(n)})^*, \tag{85}$$

with $X - X_1^{(n)}$ and $(X_1^{(n)})^*$ independent.

It is obvious that, with probability 1, $X_1^{(n)} \geq 0$, since (84) gives $(\mathbb{P}(X_1^{(n)} < 0))^n \leq \mathbb{P}(X < 0) = 0$. Next, $\mathbb{E}|X_1^{(n)}| = \mathbb{E}X_1^{(n)} = a/n \to 0$ as $n \to \infty$ implies that $X_1^{(n)} \to 0$ in $L_1$ and hence in probability. Hence $X - X_1^{(n)}$ converges in distribution to $X$.

Next, the family of random variables $(X_1^{(n)})^*$ is *tight*, because given $\epsilon > 0$, there is a $K$ such that $\mathbb{P}(X^* > K) < \epsilon$, and by (85), for all $n$, $\mathbb{P}((X_1^{(n)})^* > K) \leq \mathbb{P}(X^* > K)$. Thus, by Helly's theorem, there exists a subsequence $n_k$ of the $n$'s along which $(X_1^{(n)})^*$ converges in distribution, say $(X_1^{(n_k)})^* \Rightarrow Y$. As $n \to \infty$ along this subsequence, the pair $(X - X_1^n, (X_1^n)^*)$ converges jointly to the pair $(X, Y)$ with $X$ and $Y$ independent. From $X^* =^d (X - X_1^{(n_k)}) + (X_1^{(n_k)})^* \Rightarrow X + Y$ as $k \to \infty$ we conclude that $X^* =^d X + Y$, with $Y \geq 0$, and $X, Y$ independent. This completes the proof that if $X$ is infinitely divisible, then it satisfies (83).

That the law of $Y$ in (84) is *unique* requires a little work; we will need to know that the characteristic function $\phi$ for $X$ satisfies $\phi(u) \neq 0$ for all real $u$. Once we have this, uniqueness is easy: from (10) and (83), writing $\phi_Y$ for the characteristic function of $Y$, we have two expressions for $\phi_{X^*}(u)$, hence

$$\frac{1}{i\,\mathbb{E}X}\ \phi'(u) = \phi(u)\ \phi_Y(u). \tag{86}$$

This determines $\phi_Y(u)$, provided we know that $\phi(u) \neq 0$.

The characteristic function of *any* infinitely divisible $X$ has $\phi(u) \neq 0$ for all $u$: Feller [38, p. 500 and pp. 555–557], and Chung [30, Theorem 7.6.1], give straightforward proofs. However, under the hypothesis that (83) holds and $\mathbb{E}X$ is finite, there is a simpler proof, as follows. *Suppose that* $\phi(u) \neq 0$ *for all* $u \in (-t, t)$, *for some* $t > 0$. *From equation* (86), *for* $u \in (-t, t)$

$$(\log \phi(u))' = \frac{\phi'(u)}{\phi(u)} = i\,\mathbb{E}X\ \phi_Y(u), \ \textit{hence}\ |(\log \phi(u))'| \leq \mathbb{E}X.$$

*Since* $\phi$ *is continuous with* $\log \phi(0) = 0$, *it follows that for all* $u \in [-t, t]$, $|\log \phi(u)| \leq t\mathbb{E}X < \infty$. *If it were the case that* $\phi(u) = 0$ *for any* $u$, *we could take* $t = \inf\{|u| : \phi(u) = 0\} < \infty$ *to get a contradiction.*[18]

Finally, we prove the converse statement, that (83) implies infinite divisibility. Starting with the assumption (83), we have (86), which — with details given in the next section — lets us solve for $(\log \phi(u))'$, and integrate, to get (88) below. That (88) is the characteristic function of an infinitely divisible distribution is well-known, but to review, for the sake of a self-contained proof: the function in (88) can be expressed as the limit of characteristic functions of random variables with compound Poisson distribution, as in (36), and scaling all the Poisson parameters down by a factor of $n$, and then taking the limit, we get the distribution for the $n^{\text{th}}$ convolutional root $X_1^{(n)}$ for use in (84). □

### *11.2. The Lévy representation*

We continue to work with an $X \geq 0$ with $a := \mathbb{E}X \in (0, \infty)$, assuming also that $X$ is infinitely divisible, or equivalently, that $X$ satisfies (83). Using (86),

$$(\log \phi(u))' = \frac{\phi'(u)}{\phi(u)} = a\, i\ \phi_Y(u) \tag{87}$$

and since $\phi(0) = 1$ with $\log \phi(0) = 0$, we get

$$\log \phi(u) = a\, i \int_{t=0}^{u} \phi_Y(t)\ dt.$$

Let $\alpha$ be the distribution of $Y$ in (84), so $\alpha$ is a probability measure on $[0, \infty)$. We have

$$\int_{t=0}^{u} \phi_Y(t)\ dt = \int_{t=0}^{u} \int_y e^{ity} \alpha(dy)\ dt = \int_y \int_{t=0}^{u} e^{ity}\ dt\ \alpha(dy)$$

with the interchange justified by Fubini. We have

$$\int_{t=0}^{u} e^{ity}\ dt = \begin{cases} (e^{iuy} - 1)/(iy) & \text{if } y > 0 \\ u & \text{if } y = 0 \end{cases}$$

[18] As to the validity of taking logarithm, log continues uniquely along paths avoiding zero; see, e.g., [38, pp. 554–5], and [30, p. xv line -7 and Thm. 7.6.2].

Combining the three previous displayed equations, the characteristic function $\phi$ for $X$ may be expressed as

$$\phi(u) = \exp\left(a\left(iu\,\alpha(\{0\}) + \int_{(0,\infty)} \frac{e^{iuy}-1}{y}\,\alpha(dy)\right)\right). \tag{88}$$

To review, $a \in (0,\infty)$, $\alpha$ is the probability distribution of a nonnegative random variable $Y$, and $\phi(u)$ is the characteristic function of a random variable $X$, with $a = \mathbb{E}X$, and, with $X, Y$ independent, $X^* =^d X+Y$. We have derived (88) under the assumption that (83) holds. However, given $a \in (0,\infty)$, and a probability distribution for a nonnegative random variable $Y$, it can be seen that (88) is the characteristic function of a random variable $X$, by taking distributional limits of the discrete compound Poisson sums in (36). Then, working back through (87), one sees easily that $\mathbb{E}X = a$ and, with $X, Y$ independent, $X^* =^d X + Y$.

The calculation above, combined with Theorem 11.1, is summarized in the next theorem.

**Theorem 11.2.** *Suppose $X$ can be size biased, i.e., $X \geq 0$ and $a := \mathbb{E}X \in (0,\infty)$. If $X$ is infinitely divisible, then there exists a distribution for $Y$ such that*

$$X^* =^d X + Y, \quad \text{and } X, Y \text{ are independent.}$$

*Conversely, given that $X$ can be size biased, and that (83) holds for some $Y$, then $X$ is infinitely divisible.*

*In either case, the distribution of $Y$ is unique, $\mathbb{P}(Y \geq 0) = 1$, and $X$ has characteristic function given by (88).*

**Corollary 11.3.** *If $X$ is a nonnegative random variable with $\lambda := \mathbb{E}X \in (0,\infty)$, and $X^* =^d X + 1$, then $X$ is Poisson($\lambda$).*

A natural way to rewrite (88), motivated perhaps by the two expressions in (36), is to absorb the $1/y$ into the measure $\alpha(dy)$. Writing $\alpha_0$ for the constant $\alpha(\{0\}) = \mathbb{P}(Y = 0)$ in (88), this gives

$$\phi(u) = \exp\left(a\left(iu\alpha_0 + \int_{(0,\infty)} \left(e^{iuy}-1\right)\,\gamma(dy)\right)\right). \tag{89}$$

Here $\gamma$ is a nonnegative measure on $(0,\infty)$, with $\gamma(dy)/\alpha(dy) = 1/y$, and this allows a *broader* class than (88). To get $\mathbb{E}X < \infty$, there is the additional requirement that $\int_{(0,\infty)} y\,\gamma(dy) < \infty$ — this is the price one pays for being able to size bias. Regardless of whether $\mathbb{E}X = \infty$ or $\mathbb{E}X < \infty$, the nonnegative measure $\gamma$ can have infinite mass, due to mass near zero, and the requirement, to get a nonnegative infinitely divisible $X$, allowing $\mathbb{E}X = \infty$, is that $\int_{(0,\infty)}(1\wedge y)\gamma(dy) < \infty$. Examples 11.12 and 11.13 illustrate this, where, in both cases, $\alpha$ is a uniform distribution on an interval, and $\mathbb{E}X < \infty$.

We read (89) as: the random variable $X$ is the constant $a\,\alpha_0$, plus the sum of arrivals, in the Poisson process on $(0,\infty)$ with intensity measure $a\,\gamma$. Formula

(89)[19] is called the Lévy–Khintchine formula in the survey paper on subordinators [20], the one difference being that the random variable $X$ representing the value of the subordinator at time $a$ is also allowed to have $\mathbb{P}(X = \infty) = 1 - \exp(-ka) > 0$, where $k$ is called the *killing rate*.[20]

### *11.3. The size bias equation*

When $X, Y$ are both discrete or both absolutely continuous, it is worth highlighting how (4), together with (83), yields a simple relation satisfied by the mass functions or densities. Sato [78] Section 51, especially Corollary 51.2, already highlights these relations, though of course without referring to them as being size bias relations.

In the discrete case, if (83) holds, then $f_{X^*}$ is the convolution of $f_X$ and $f_Y$: $f_{X^*}(x) = \sum_y f_X(x-y) f_Y(y)$, and combining with (4) yields, for all $x > 0$,

$$f_X(x) = \frac{a}{x} \sum_y f_X(x-y) f_Y(y). \tag{90}$$

A common special case is that $Y$ is supported on the positive integers, and $X$ on the nonnegative integers, so that considering $f_Y$ as known, and $f_X$ to be found, the homogeneous system (90) specifies a recursion: starting from $f_X(0) = c$, for $m = 0, 1, 2, \ldots$,

$$f_X(m+1) = \frac{a}{m+1} \sum_{0 \le i \le m} f_X(i) f_Y(m+1-i), \tag{91}$$

and the initial value $c$ is determined by $1 = \sum_{i \ge 0} f_X(i)$. Furthermore, from (36) and (88) we know that $X = \sum_{i \ge 1} i Z_i$ with $Z_i$ independent Poisson$(\lambda_i)$, $\lambda := \sum \lambda_i < \infty$, $f_Y(i) = i\lambda_i / a$, hence $f_X(0) = P(Z_1 = Z_2 = \cdots = 0) = e^{-\lambda}$. The relation (91) was used in [10], where it was referred to as a result from [74]. The situation with $X = \sum_{1 \le i \le n} i Z_i$ with $Z_i$ independent Poisson$(\lambda_i)$ is universal to *combinatorial assemblies*; here $X$ is usually denoted as $T_n$, and conditional on the event $(T_n = n)$ one has a labelled combinatorial object of total size $n$, in which there are $Z_i$ components of size $i$, jointly for $i = 1$ to $n$. See [11, 7].

Likewise, in the absolutely continuous case, where $X$ and $Y$ have densities, if (83) holds, then $f_{X^*}$ is the convolution of $f_X$ and $f_Y$: $f_{X^*}(x) = \int_y f_X(x-y) f_Y(y)\, dy$. Combined with (4), this says that for all $x > 0$,

$$f_X(x) = \frac{a}{x} \int_y f_X(x-y) f_Y(y)\, dy. \tag{92}$$

[19] Or any of its cousins, such as the Laplace transform version — since the characteristic function $\phi(\cdot)$, moment generating function $M(\cdot)$, and Laplace transform $L(\cdot)$ are essentially the same, in detail $\phi(u) := \mathbb{E}e^{iuX}$, $M(\beta) := \mathbb{E}e^{\beta X}$, $L(t) := \mathbb{E}e^{-tX}$, allowing the formal substitutions $iu = \beta = -t$.

[20] And when there is killing, then the Laplace transform *is preferable* to the characteristic function; see the previous footnote.

### *11.4. Examples of infinitely divisible distributions for nonnegative random variables*

Of course, the Lévy representation (89) yields *all* examples of nonnegative infinitely divisible distributions. However, *recognizing* when a given distribution for $X$ takes the form (88) or (89) remains a nontrivial problem. We present our favorite examples in which Theorem 11.1 provides a convenient criterion, and we will use the notation from Theorem 11.1, in particular (83).

#### *11.4.1. Discrete examples*

**Example 11.4.** $\mathbb{P}(Y = 1) = 1$*; $X$ is Poisson(a).*

**Example 11.5.** *For $p \in (0,1)$, $\mathbb{P}(Y = k) = (1-p)^k/k$. When $a = (1-p)/p$, $X$ is geometric, with $\mathbb{P}(X = n) = (1-p)^n p, n \geq 0$. When $ap/(1-p)$ is a positive integer, $X$ is negative binomial.*

*The infinite divisibility of geometric and negative binomial distributions plays a key role in estimates comparing logarithmic combinatorial structures with their limits; see [7]. The compound Poisson representation of the geometric is the starting point for a coupling, in [4], showing that a random integer may be chosen uniformly from 1 to $n$, on the same probability space with a Poisson-Dirichlet process $(L_1, L_2, \ldots)$, so that if $P_i$ is the $i^{\text{th}}$ largest factor of the random integer,*[21] *then $\mathbb{E}\sum_{i\geq 1} |\log P_i - (\log n)L_i| = O(\log\log n)$. This construction is analogous to Skorohod embedding: it starts with the continuum limit process – Poisson-Dirichlet instead of Brownian motion– and constructs the nearby (in the limit) discrete random object –the random integer expressed as a product of primes instead of a random walk– as a deterministic function of the continuum limit process, together with a small amount of auxiliary randomization.*

A necessary and sufficient condition for a nonnegative integer valued random variable to be infinitely divisible is given in [56], and a useful *sufficient* condition is given in [89]. The sufficient condition is *log-convexity*: the support of $X$ is the nonnegative integers, and for all $n \geq 1$, $\mathbb{P}(X = n-1)\mathbb{P}(X = n+1) \geq \mathbb{P}(X = n)^2$. Example 11.4 shows that the sufficient condition of log-convexity is not necessary — any Poisson distribution is log-*concave*, rather than log-*convex*. See [9] for a discussion of how the sufficiency of log-convexity is perhaps attributable to Kaluza, [54]. Of course, for any constant $c$, $X$ is infinitely divisible if, and only if, $c + X$ is infinitely divisible; this remark is often used with $c = \pm 1$. There are several famous discrete distributions that can be seen to be infinitely divisible via log-convexity; some examples of this type are given in [89], and two of our favorite examples are the following:

**Example 11.6.** *The zeta distributions: For $s > 0$, $\mathbb{P}(X = n) = n^{-s}/\zeta(s)$, $n \geq 1$.*

**Example 11.7.** *The simplest power law, $\mathbb{P}(X \geq n) = 1/n$ for $n \geq 1$.*

[21] with the convention that $P_i = 1$ when $i$ exceeds the number of prime factors, including multiplicities

*11.4.2. Continuous examples*

**Example 11.8.** *$Y$ is exponential, with $\mathbb{P}(Y > t) = e^{-t}$ for $t \geq 0$. When $a = 1$, $X =^d Y$, and $X^* =^d X + Y$ is the sum of two independent copies of $X$, as observed in Section 3 on the waiting time paradox. For positive integers $a$, $X$ is the time of the $a^{\text{th}}$ arrival in a standard Poisson process. For general $a > 0$, $X$ has the Gamma distribution, with shape parameter $a$.*

*In the Lévy representation (89) for the characteristic function of the Gamma random variable $X$, we have $\gamma(dy) = e^{-y}/y\, dy$. This measure $\gamma$, or the increasing process it governs, is also known as the Moran subordinator, and used to construct the Poisson-Dirchlet process; see [58].*

**Example 11.9.** *Pareto distributions, of the form $\mathbb{P}(X > t) = (1+t)^{-\alpha}$, $\alpha > 0$.*

*This is the example for which Thorin [87] first developed his theory of generalized Gamma convolutions, which is a subclass of the infinitely divisible distributions for positive random variables. See [22], as well as [23].*

**Example 11.10.** *The lognormal distributions. Again, this is from Thorin in 1977, [86], and his proof is based on a generalized Gamma convolution.*

**Example 11.11.** *Distributions with a log-convex density.*

*Taking limits of discrete distributions on the nonnegative integers with log-convex pointmass function, Sato [78, Theorem 5.1.4] shows that if $X$ has a density $f$ on $(0, \infty)$, such that $\log f$ is convex on $(0, \infty)$, then $X$ is infinitely divisible. This also shows that the Pareto distributions are infinitely divisible!*

The next two examples, Examples 11.12 and 11.13, arise by taking $Y$ in (83) to be uniformly distributed on a bounded interval of nonnegative numbers. Up to scaling, *any* such $Y$ is either uniformly distributed on $(0, 1)$, or else on $(b, 1)$ for some $0 < b < 1$. In the former case, $X$ has an absolutely continuous distribution, and the latter case the distribution of $X$ has an atom and an absolutely continuous part.

**Example 11.12.** *$Y$ is uniform $(0, 1)$, leading to Dickman's function $\rho$, and its convolution powers.*

*In (83), take $Y$ to be the standard uniform random variable on $(0, 1)$. Then (88) specializes to*

$$\phi_X(u) = \exp\left( a \int_0^1 \frac{e^{iuy} - 1}{y}\, dy \right), \tag{93}$$

*and (92) specializes to*

$$f_X(x) = \frac{a}{x} \int_{y=0}^{1} f_X(x - y)\, dy = \frac{a}{x} \int_{x-1}^{x} f_X(z)\, dz. \tag{94}$$

*Here as always, $a = \mathbb{E}X$; the choice $a = 1$ yields $f_X(x) = e^{-\gamma}\rho(x)$, where $\rho$ is Dickman's function, of central importance in the study of integers without large prime factors; see [85] and [7, Section 4.2]. For the general case $a \in (0, \infty)$, the*

*density $f_X$ is a "convolution power of Dickman's function," normalized to be a probability density; see [47] .*

**Example 11.13.** *$Y$ is uniform $(b,1)$ for $0 < b < 1$, leading to Buchstab's function $\omega$, and the limit probability for* logarithmic *structures to have all parts in a range excluding small parts, or both small and large parts.*

*Now (89) becomes*

$$\phi_X(u) = \exp\left(\frac{a}{1-b}\int_b^1 \frac{e^{iuy}-1}{y}\,dy\right), \tag{95}$$

*with $0 < b < 1$. Unlike Example 11.12, $X$ is no longer absolutely continuous, since $\mathbb{P}(X=0) = b^{a/(1-b)} > 0$.*

*This computation of $\mathbb{P}(X=0)$ is easy to understand, by viewing (89) as the specification that $X$ is the sum of the arrivals in the Poisson process with arrival intensity measure $a\,\gamma$, where $a\gamma(dy) = a/(1-b)\ 1(b<y<1)\ dy/y$. The expected number of arrivals in this Poisson process is $\lambda = \int_b^1 a/(1-b)\ dy/y$, and of course $\mathbb{P}(X=0) = e^{-\lambda}$. See [6].*

*The size bias equation, which was (94) for the case $b=0$, is more complicated with $0 < b < 1$: the distribution of $X$ has pointmass $b^{a/(1-b)} > 0$, and a defective density $f_X$ whose support is $\cup_{k\geq 1}[kb,k]$. The size bias equation obtained by combining (4) with (83) takes the form: for $x > 0$,*

$$f_X(x) = \frac{a}{x}\left(b^{a/(1-b)}\,\frac{1(b<x<1)}{1-b} + \int_{y=b}^1 \frac{f_X(x-y)}{1-b}\,dy\right). \tag{96}$$

*We briefly explain the natural importance of Example 11.13. Let $f_X^{(b)}$ be the density of $X$, for $0 < b < 1$ and $a = 1-b$. This density arises in the study of random permutations; see [7, Section 4.3]. Directly, $f_X^{(b)}(1)$ governs the asymptotic probability that a random permutation of $n$ objects has only cycles of length at least $bn$. Scale invariance also leads, for fixed $b \in (0,1)$, to $f_X^{(b)}(u)$ governing the asymptotic probability that a random permutation on $n$ objects has only cycles with lengths in $(bn/u, n/u)$, for any $u > 1$. Scale invariance also leads to $\omega(u) = f_X^{(1/u)}(1)$, with Buchstab's function $\omega$ governing integers free of small prime factors; see [85].*

## References


[1] Romain Abraham and Jean-François Delmas. Local limits of conditioned Galton-Watson trees: the infinite spine case. *Electron. J. Probab.*, 19:no. 2, 19, 2014. MR3164755

[2] David Aldous. Tree-valued Markov chains and Poisson-Galton-Watson distributions. In *Microsurveys in discrete probability (Princeton, NJ, 1997)*, volume 41 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 1–20. Amer. Math. Soc., Providence, RI, 1998. MR1630406

[3] David Applebaum. Lévy processes and stochastic calculus, 2009. MR2512800

[4] R. Arratia. On the amount of dependence in the prime factorization of a uniform random integer. In *Contemporary combinatorics*, volume 10 of *Bolyai Soc. Math. Stud.*, pages 29–91. János Bolyai Math. Soc., Budapest, 2002. MR1919568

[5] R. Arratia, S. Garibaldi, and J. Killian. Asymptotic distribution for the birthday problem with multiple coincidences, via an embedding of the collision process. *Random Structures and Algorithms*, 2015. MR3481270

[6] Richard Arratia. On the central role of scale invariant Poisson processes on $(0, \infty)$. In *Microsurveys in discrete probability (Princeton, NJ, 1997)*, volume 41 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 21–41. Amer. Math. Soc., Providence, RI, 1998. MR1630407

[7] Richard Arratia, A. D. Barbour, and Simon Tavaré. *Logarithmic combinatorial structures: a probabilistic approach*. EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich, 2003. MR2032426

[8] Richard Arratia and Peter Baxendale. Bounded size bias coupling: a Gamma function bound, and universal Dickman-function behavior. *Probability Theory and Related Fields*, pages 1–19, 2014. MR3383333

[9] Richard Arratia, Thomas M. Liggett, and Malcolm J. Williamson. Scale-free and power law distributions via fixed points and convergence of (thinning and conditioning) transformations. *Electron. Commun. Probab.*, 19:no. 39, 10, 2014. MR3225870

[10] Richard Arratia and Simon Tavaré. The cycle structure of random permutations. *Ann. Probab.*, 20(3):1567–1591, 1992. MR1175278

[11] Richard Arratia and Simon Tavaré. Independent process approximations for random combinatorial structures. *Advances in Mathematics*, 104:90–154, 1994. MR1272071

[12] Krishna B. Athreya and Peter E. Ney. *Branching processes*. Springer-Verlag, New York-Heidelberg, 1972. Die Grundlehren der mathematischen Wissenschaften, Band 196. MR0373040

[13] P. Baldi, Y. Rinott, and C. Stein. A normal approximation for the number of local maxima of a random function on a graph. In *Probability, statistics, and mathematics*, pages 59–81. Academic Press, Boston, MA, 1989. MR1031278

[14] Pierre Baldi and Yosef Rinott. On normal approximations of distributions in terms of dependency graphs. *Ann. Probab.*, 17(4):1646–1650, 1989. MR1048950

[15] A. D. Barbour, Lars Holst, and Svante Janson. *Poisson approximation*, volume 2 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1992. Oxford Science Publications. MR1163825

[16] J. Bartroff, L. Goldstein, and Ü Işlak. Bounded size biased couplings for log concave distributions and concentration of measure for occupancy models. Preprint. 2013. MR3788174

[17] Christian Berg. From discrete to absolutely continuous solutions of indeterminate moment problems. *Arab J. Math. Sci.*, 4(2):1–18, 1998. MR1667218

[18] Christian Berg. On some indeterminate moment problems for measures on a geometric progression. *J. Comput. Appl. Math.*, 99(1-2):67–75, 1998. MR1662684

[19] Jean Bertoin. *Lévy processes*, volume 121 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1996. MR1406564

[20] Jean Bertoin. Subordinators: examples and applications. In *Lectures on probability theory and statistics (Saint-Flour, 1997)*, volume 1717 of *Lecture Notes in Math.*, pages 1–91. Springer, Berlin, 1999. MR1746300

[21] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication. MR1324786

[22] L. Bondesson. Generalized gamma convolutions and complete monotonicity. *Probab. Theory Related Fields*, 85(2):181–194, 1990. MR1050743

[23] Lennart Bondesson, Jan Grandell, and Jaak Peetre. The life and work of Olof Thorin (1912–2004). *Proc. Est. Acad. Sci.*, 57(1):18–25, 2008. MR2555072

[24] Mark Brown. Exploiting the waiting time paradox: applications of the size-biasing transformation. *Probab. Engrg. Inform. Sci.*, 20(2):195–230, 2006. MR2261286

[25] Louis H. Y. Chen. Poisson approximation for dependent trials. *Ann. Probability*, 3(3):534–545, 1975. MR0428387

[26] Louis H. Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Probability and its Applications (New York). Springer, Heidelberg, 2011. MR2732624

[27] T. S. Chihara. A characterization and a class of distribution functions for the Stieltjes-Wigert polynomials. *Canad. Math. Bull.*, 13:529–532, 1970. MR0280761

[28] Yuan Shih Chow and Henry Teicher. *Probability theory*. Springer Texts in Statistics. Springer-Verlag, New York, third edition, 1997. Independence, interchangeability, martingales. MR1476912

[29] Jacob Stordal Christiansen. The moment problem associated with the Stieltjes-Wigert polynomials. *J. Math. Anal. Appl.*, 277(1):218–245, 2003. MR1954473

[30] Kai Lai Chung. *A course in probability theory*. Academic Press Inc., San Diego, CA, third edition, 2001. MR1796326

[31] William G. Cochran. *Sampling techniques*. John Wiley & Sons, New York-London-Sydney, third edition, 1977. Wiley Series in Probability and Mathematical Statistics. MR0474575

[32] Nicholas Cook, Larry Goldstein, and Tobias Johnson. Size biased couplings and the spectral gap for random regular graphs. *Ann. Probab.*, 46(1):72–125, 2018. MR3758727

[33] David Cox. *Selected statistical papers of Sir David Cox. Vol. I*. Cambridge University Press, Cambridge, 2005. Design of investigations, statistical methods and applications, Edited by D. J. Hand and A. M. Herzberg, containing the article "Some sampling problems in technology". MR2238853

[34] D. J. Daley and D. Vere-Jones. *An introduction to the theory of*

*point processes. Vol. II.* Probability and its Applications (New York). Springer, New York, second edition, 2008. General theory and structure. MR2371524

[35] Donald Dawson. *Introductory Lectures on Stochastic Population Systems.* 2017. From arxiv.org/abs/1705.03781.

[36] J. L. Doob. Renewal theory from the point of view of the theory of probability. *Trans. Amer. Math. Soc.*, 63:422–438, 1948. MR0025098

[37] Rick Durrett. *Probability: theory and examples.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010. MR2722836

[38] William Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York, 1971. MR0270403

[39] Subhankar Ghosh and Larry Goldstein. Applications of size biased couplings for concentration of measures. *Electron. Commun. Probab.*, 16:70–83, 2011. MR2763529

[40] Subhankar Ghosh and Larry Goldstein. Concentration of measures via size-biased couplings. *Probab. Theory Related Fields*, 149(1-2):271–278, 2011. MR2773032

[41] Larry Goldstein and Mathew D. Penrose. Normal approximation for coverage models over binomial point processes. *Ann. Appl. Probab.*, 20(2):696–721, 2010. MR2650046

[42] Larry Goldstein and Gesine Reinert. Total variation distance for Poisson subset numbers. *Ann. Comb.*, 10(3):333–341, 2006. MR2284274

[43] Larry Goldstein and Yosef Rinott. Multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Probab.*, 33(1):1–17, 1996. MR1371949

[44] Louis Gordon. Estimation for large successive samples with unknown inclusion probabilities. *Adv. in Appl. Math.*, 14(1):89–122, 1993. MR1204057

[45] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and random processes.* Oxford University Press, New York, third edition, 2001. MR2059709

[46] Morris H. Hansen and William N. Hurwitz. On the theory of sampling from finite populations. *Ann. Math. Statistics*, 14:333–362, 1943. MR0009832

[47] Douglas Hensley. The convolution powers of the Dickman function. *J. London Math. Soc. (2)*, 33(3):395–406, 1986. MR0850955

[48] C. C. Heyde. On a property of the lognormal distribution. *J. Roy. Statist. Soc. Ser. B*, 25:392–393, 1963. MR0171336

[49] Peter Jagers. On Palm probabilities. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 26:17–32, 1973. MR0339330

[50] Karen Kafadar and Philip C. Prorok. Effect of length biased sampling of unobserved sojourn times on the survival distribution when disease is screen detected. *Stat. Med.*, 28(16):2116–2146, 2009. MR2751510

[51] Olav Kallenberg. Characterization and convergence of random measures and point processes. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 27:9–21, 1973. MR0431374

[52] Olav Kallenberg. *Random measures*. Akademie-Verlag, Berlin, 1975. Schriftenreihe des Zentralinstituts für Mathematik und Mechanik bei der Akademie der Wissenschaften der DDR, Heft 23. MR0431372

[53] Olav Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002. MR1876169

[54] Th. Kaluza. Über die Koeffizienten reziproker Potenzreihen. *Math. Z.*, 28(1):161–170, 1928. MR1544949

[55] Edward L. Kaplan. Transformations of stationary random sequences. *Math. Scand.*, 3:127–149, 1955. MR0072381

[56] S. K. Katti. Infinite divisibility of integer-valued random variables. *Ann. Math. Statist.*, 38:1306–1308, 1967. MR0215333

[57] Harry Kesten. Subdiffusive behavior of random walk on a random cluster. *Ann. Inst. H. Poincaré Probab. Statist.*, 22(4):425–487, 1986. MR0871905

[58] J. F. C. Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1993. Oxford Science Publications. MR1207584

[59] R. Leipnik. The lognormal distribution and strong nonuniqueness of the moment problem. *Teor. Veroyatnost. i Primenen.*, 26(4):863–865, 1981. Also appeared in J. Prob. Appl. 863–865, 1981. MR0636784

[60] Roy B. Leipnik. On lognormal random variables. I. The characteristic function. *J. Austral. Math. Soc. Ser. B*, 32(3):327–347, 1991. MR1079462

[61] Michel Loève. *Probability theory*. Third edition. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London, 1963. MR0203748

[62] Marcos López-García. Characterization of distributions with the length-bias scaling property. *Electron. Commun. Probab.*, 14:186–191, 2009. MR2505174

[63] Ho Ming Luk. *Stein's method for the Gamma distribution and related statistical applications*. ProQuest LLC, Ann Arbor, MI, 1994. Thesis (Ph.D.)–University of Southern California. MR2693204

[64] Russell Lyons, Robin Pemantle, and Yuval Peres. Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Ann. Probab.*, 23(3):1125–1138, 1995. MR1349164

[65] Russell Lyons and Yuval Peres. *Probability on trees and networks*, volume 42 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York, 2016. MR3616205

[66] Hiroshi Midzuno. On the sampling system with probability proportionate to sum of sizes. *Ann. Inst. Statist. Math., Tokyo*, 3:99–107, 1952. MR0050840

[67] Jan Obłój. The Skorokhod embedding problem and its offspring. *Probab. Surv.*, 1:321–390, 2004. MR2068476

[68] Anthony G. Pakes. Length biasing and laws equivalent to the log-normal. *J. Math. Anal. Appl.*, 197(3):825–854, 1996. MR1373083

[69] Anthony G. Pakes and Ravindra Khattree. Length-biasing, characterizations of laws and the moment problem. *Austral. J. Statist.*, 34(2):307–322, 1992. MR1193781

[70] Erol Peköz and Adrian Röllin. Exponential approximation for the nearly critical Galton-Watson process and occupation times of Markov chains. *Electron. J. Probab.*, 16:no. 51, 1381–1393, 2011. MR2827464

[71] Jim Pitman and Nathan Ross. Archimedes, Gauss, and Stein. *Notices Amer. Math. Soc.*, 59(10):1416–1421, 2012. MR3025901

[72] Jim Pitman and Marc Yor. Infinitely divisible laws associated with hyperbolic functions. *Canad. J. Math.*, 55(2):292–330, 2003. MR1969794

[73] Pomegranate Apps. *MathStudio Version 5.4*. Pomegranate Apps, Minneapolis MN, 2013.

[74] Mohsen Pourahmadi. Taylor expansion of $\exp(\sum_{k=0}^{\infty} a_k z^k)$ and some applications. *Amer. Math. Monthly*, 91(5):303–307, 1984. MR0740245

[75] T. J. Rao. On the variance of the ratio estimator for Midzuno-Sen sampling scheme. *Metrika*, 10:89–91, 1966. MR0195223

[76] L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 1.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, second edition, 1994. Foundations. MR1331599

[77] Nathan Ross. Fundamentals of Stein's method. *Probab. Surv.*, 8:210–293, 2011. MR2861132

[78] Ken-iti Sato. *Lévy processes and infinitely divisible distributions*, volume 68 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. Translated from the 1990 Japanese original, Revised by the author. MR1739520

[79] Zhan Shi. *Branching random walks*, volume 2151 of *Lecture Notes in Mathematics*. Springer, Cham, 2015. Lecture notes from the 42nd Probability Summer School held in Saint Flour, 2012, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR3444654

[80] F. W. Steutel. *Preservation of infinite divisibility under mixing and related topics.*, volume 33 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1970. MR0278355

[81] F. W. Steutel. Some recent results in infinite divisibility. *Stochastic Processes Appl.*, 1:125–143, 1973. MR0372948

[82] Fred W. Steutel and Klaas van Harn. *Infinite divisibility of probability distributions on the real line*, volume 259 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker Inc., New York, 2004. MR2011862

[83] T.-J. Stieltjes. Recherches sur les fractions continues. *Ann. Fac. Sci. Toulouse Sci. Math. Sci. Phys.*, 8(4):J1–J122, 1894. MR1508159

[84] Thomas Jan Stieltjes. *Œuvres complètes/Collected papers. Vol. I, II.* Springer-Verlag, Berlin, 1993. Reprint of the 1914–1918 edition, Edited and with a preface and a biographical note by Gerrit van Dijk, With additional biographical and historical material by Walter Van Assche, Frits Beukers, Wilhelmus A. J. Luxemburg and Herman J. J. te Riele. MR1272017

[85] Gérald Tenenbaum. *Introduction to analytic and probabilistic number theory*, volume 46 of *Cambridge Studies in Advanced Mathematics*. Cambridge

University Press, Cambridge, 1995. Translated from the second French edition (1995) by C. B. Thomas. MR1342300

[86] Olof Thorin. On the infinite divisibility of the lognormal distribution. *Scand. Actuar. J.*, (3):121–148, 1977. MR0552135

[87] Olof Thorin. On the infinite divisibility of the Pareto distribution. *Scand. Actuar. J.*, (1):31–40, 1977. MR0431333

[88] Hermann Thorisson. *Coupling, stationarity, and regeneration*. Probability and its Applications (New York). Springer-Verlag, New York, 2000. MR1741181

[89] W. D. Warde and S. K. Katti. Infinite divisibility of discrete distributions. II. *Ann. Math. Statist.*, 42:1088–1090, 1971. MR0293691