J. Eur. Math. Soc. (Online first) DOI 10.4171/JEMS/1403

Jean Bertoin

Counterbalancing steps at random in a random walk

Received November 27, 2020; revised May 19, 2022

Abstract. A random walk with counterbalanced steps is a process of partial sums $\check{S}(n) = \check{X}_1 + \cdots + \check{X}_n$ whose steps $\check{X}_n$ are given recursively as follows. For each $n \geq 2$, with a fixed probability p, $\check{X}_n$ is a new independent sample from some fixed law μ, and with complementary probability $1-p$, $\check{X}_n = -\check{X}_{v(n)}$ counterbalances a previous step, with $v(n)$ a uniform random pick from $\{1, \ldots, n-1\}$. We determine the asymptotic behavior of $\check{S}(n)$ in terms of p and the first two moments of μ. Our approach relies on a coupling with a reinforcement algorithm due to H. A. Simon, and on properties of random recursive trees and Eulerian numbers, which may be of independent interest. The method can be adapted to the situation where the step distribution μ belongs to the domain of attraction of a stable law.

Keywords. Reinforcement, random walk, random recursive tree, Eulerian numbers, Yule–Simon model

1. Introduction

In short, the purpose of the present work is to investigate long time effects of an algorithm for counterbalancing steps in a random walk. As we shall first explain, our motivation stems from a nearest neighbor process on the integer lattice, known as the elephant random walk.

The elephant random walk is a stochastic process with memory on $\mathbb{Z}$, which records the trajectory of an elephant that makes steps with unit length left or right at each positive integer time. It has been introduced by Schütz and Trimper [31] and triggered a growing interest in the recent years; see, for instance, [4,5,15,16,22,23], and also [2,6,7,14,20,26] for related models. The dynamics depend on a parameter $q \in [0, 1]$ and can be described as follows. Let us assume that the first step of the elephant is a Rademacher variable, that is, equals $+1$ or -1 with probability $1/2$. For each time $n \geq 2$, the elephant remembers a step picked uniformly at random among those it made previously, and decides either to repeat it with probability q, or to make the opposite step with complementary probability

Jean Bertoin: Institute of Mathematics, University of Zürich, 8057 Zürich, Switzerland;
jean.bertoin@math.uzh.ch

Mathematics Subject Classification (2020): Primary 05C05; Secondary 60F05, 60G50, 05A05

$1-q$. Obviously, each step of the elephant then has the Rademacher law, although the sequence of steps is clearly not stationary.

Roughly speaking, it seems natural to generalize these dynamics and allow steps to have an arbitrary distribution on $\mathbb{R}$, say μ. In this direction, Kürsten [23] pointed to an equivalent way of describing the dynamics of the elephant random walk which makes such generalization non-trivial.[1] Let $p \in [0,1]$, and imagine a walker who makes at each time a step which is either, with probability p, a new independent random variable with law μ, or, with probability $1-p$, a repetition of one of his preceding steps picked uniformly at random. It is immediately checked that when μ is the Rademacher distribution, then the walker follows the same dynamics as the elephant random walk with parameter $q = 1 - p/2$. When μ is an isotropic stable law, this is the model referred to as the shark random swim by Businger [14], and more generally, when μ is arbitrary, this is the step reinforced random walk that has been studied lately in e.g. [8–11].

The model of Kürsten yields an elephant random walk only with parameter $q \in [1/2, 1]$; nonetheless the remaining range can be obtained by a simple modification. Indeed, let again $p \in [0,1]$ and imagine now a repentant walker who makes at each time a step which is either, with probability p, a new independent random variable with law μ, or, with probability $1-p$, the *opposite* of one of his previous steps picked uniformly at random. When μ is the Rademacher distribution, we simply get the dynamics of the elephant random walk with parameter $q = p/2 \in [0, 1/2]$.

More formally, we consider a sequence (X_n) of i.i.d. real random variables with some given law μ and a sequence $(\varepsilon_n)_{n\geq 2}$ of i.i.d. Bernoulli variables with parameter $p \in [0,1]$, which we assume furthermore to be independent of (X_n). We construct a counterbalanced sequence $(\check{X}_n)$ by interpreting each $\{\varepsilon_n = 0\}$ as a counterbalancing event and each $\{\varepsilon_n = 1\}$ as an innovation event. Specifically, we agree that $\varepsilon_1 = 1$ for definiteness and denote the number of innovations after n steps by

$$\mathrm{i}(n) := \sum_{j=1}^{n} \varepsilon_j \quad \text{for } n \geq 1.$$

We introduce a sequence $(v(n))_{n\geq 2}$ of independent variables, where each $v(n)$ has the uniform distribution on $\{1, \ldots, n-1\}$, and which is also independent of (X_n) and (ε_n). We then define recursively

$$\check{X}_n := \begin{cases} -\check{X}_{v(n)} & \text{if } \varepsilon_n = 0, \\ X_{\mathrm{i}(n)} & \text{if } \varepsilon_n = 1. \end{cases} \tag{1}$$

Note that the same step can be counterbalanced several times, and also that certain steps counterbalance previous steps which in turn already counterbalanced earlier ones. The

[1] Note that merely replacing the Rademacher distribution for the first step of the elephant by μ would not be interesting, as one would then just get the evolution of the elephant random walk multiplied by some random factor with law μ.

process

$$\check{S}(n) := \check{X}_1 + \cdots + \check{X}_n, \quad n \geq 0,$$

which records the positions of the repentant walker as a function of time, is called here a *random walk with counterbalanced steps*. Note that for $p = 1$, i.e. when no counterbalancing events occur, $\check{S}$ is just a usual random walk with i.i.d. steps.

In short, we are interested in understanding how counterbalancing steps affect the asymptotic behavior of random walks. We first introduce some notation. Recall that μ denotes the distribution of the first step $X_1 = \check{X}_1$ and write

$$m_k := \mathbb{E}(X_1^k) = \int_{\mathbb{R}} x^k \, \mu(\mathrm{d}x)$$

for the moment of order $k \geq 1$ of X_1 whenever $X_1 \in L^k(\mathbb{P})$. To start with, we point out that if the first moment is finite, then the algorithm (1) yields the recursive equation

$$\mathbb{E}(\check{S}(n+1)) = pm_1 + (1 - (1-p)/n)\mathbb{E}(\check{S}(n)), \quad n \geq 1,$$

with the initial condition $\mathbb{E}(\check{S}(1)) = m_1$. It follows easily that

$$\mathbb{E}(\check{S}(n)) \sim \frac{p}{2-p} m_1 n \quad \text{as } n \to \infty;$$

see e.g. [18, Lemma 4.1.2]. Our first result about the ballistic behavior should therefore not come as a surprise.

Proposition 1.1. *Let $p \in [0,1]$. If $X_1 \in L^1(\mathbb{P})$, then*

$$\lim_{n\to\infty} \frac{\check{S}(n)}{n} = \frac{p}{2-p} m_1 \quad \textit{in probability.}$$

We see in particular that counterbalancing steps reduces the asymptotic velocity of a random walk by a factor $p/(2-p) < 1$. The velocity is smaller when the innovation rate p is smaller (i.e. when counterbalancing events have a higher frequency), and vanishes as p approaches $0+$.

The main purpose of this work is to establish the asymptotic normality when μ has a finite second moment.

Theorem 1.2. *Let $p \in (0,1]$. If $X_1 \in L^2(\mathbb{P})$, then*

$$\lim_{n\to\infty} \frac{\check{S}(n) - \frac{p}{2-p} m_1 n}{\sqrt{n}} = \mathcal{N}\left(0, \frac{m_2 - \left(\frac{p}{2-p} m_1\right)^2}{3-2p}\right) \quad \textit{in distribution,}$$

where the right-hand side denotes a centered Gaussian variable parametrized by mean and variance.

It is interesting to observe that the variance of the Gaussian limit depends linearly on the square m_1^2 of the first moment and the second moment m_2 of μ only, although not

just on the variance $m_2 - m_1^2$ (except, of course, for $p = 1$). Furthermore, it is not always a monotone function[2] of the innovation rate p, and does not vanish when p tends to 0 either.

Actually, our proofs of Proposition 1.1 and Theorem 1.2 provide a much finer analysis than what is encapsulated by those general statements. Indeed, we shall identify the main actors for the evolution of $\check{S}$ and their respective contributions to its asymptotic behavior. In short, we shall see that the ballistic behavior stems from those of the variables X_j that have been used just once by the algorithm (1) (in particular, they have not yet been counterbalanced), whereas the impact of variables that occurred twice or more, regardless of their signs $\pm$, is asymptotically negligible as far as only velocity is concerned. Asymptotic normality is more delicate to analyze. We shall show that, roughly speaking, it results from the combination of, on the one hand, the central limit theorem for certain centered random walks, and on the other hand, Gaussian fluctuations for the asymptotic frequencies of some pattern induced by (1).

Our analysis relies on a natural coupling of the counterbalancing algorithm (1) with a basic linear reinforcement algorithm which was introduced a long time ago by H. A. Simon [32] to explain the occurrence of certain heavy tailed distributions in a variety of empirical data. Specifically, Simon defined recursively a sequence denoted here by $(\hat{X}_n)$ (beware of the difference of notation between $\hat{X}$ and $\check{X}$) via

$$\hat{X}_n := \begin{cases} \hat{X}_{v(n)} & \text{if } \varepsilon_n = 0, \\ X_{\mathrm{i}(n)} & \text{if } \varepsilon_n = 1. \end{cases} \tag{2}$$

We stress that the same Bernoulli variables ε_n and the same uniform variables $v(n)$ are used to run both Simon's algorithm (2) and (1); in particular either $\check{X}_n = \hat{X}_n$ or $\check{X}_n = -\hat{X}_n$. It might then seem natural to refer to (1) and (2) respectively as negative and positive reinforcement algorithms. However, in the literature, negative reinforcement usually refers to a somehow different notion (see e.g. [29]), and we shall avoid using this terminology.

A key observation is that (1) can be recovered from (2) as follows. Simon's algorithm naturally encodes a genealogical forest with set of vertices $\mathbb{N} = \{1, 2, \ldots\}$ and edges $(v(j), j)$ for all $j \geq 2$ with $\varepsilon_j = 0$; see Figure 1 in Section 3. Then $\check{X}_n = \hat{X}_n$ if the vertex n belongs to an even generation of its tree component, and $\check{X}_n = -\hat{X}_n$ if n belongs to an odd generation. On the other hand, the statistics of Simon's genealogical forest can be described in terms of independent random recursive trees (see e.g. [17, Chapter 6] for background) conditionally on their sizes. This leads us to investigate the difference $\Delta(\mathbb{T}_k)$ between the number of vertices at even generations and the number of vertices at odd generations in a random recursive tree $\mathbb{T}_k$ of size $k \geq 1$. The law of $\Delta(\mathbb{T}_k)$ can

[2]For instance, in the simplest case when μ is a Dirac point mass, i.e. $m_2 = m_1^2$, the variance is given by $\frac{4(1-p)m_2}{(3-2p)(2-p)^2}$ and reaches its maximum for $p = (9 - \sqrt{17})/8 \approx 0.6$. At the other extreme, when μ is centered, i.e. $m_1 = 0$, the variance is given by $m_2/(3-2p)$ and hence increases with p.

be expressed in terms of Eulerian numbers, and properties of the latter enable us either to compute explicitly or estimate certain quantities which are crucial for the proofs of Proposition 1.1 and Theorem 1.2.

It is interesting to compare asymptotic behaviors for counterbalanced steps with those for reinforced steps. If we write $\hat{S}(n) = \hat{X}_1 + \cdots + \hat{X}_n$ for the random walk with reinforced steps, then it is known that the law of large numbers holds for $\hat{S}$, namely $\hat{S}(n)/n \to m_1$ in L^1 when $\int_{\mathbb{R}} |x| \, \mu(\mathrm{d}x) < \infty$, independently of the innovation parameter p. Further, regarding fluctuations when $\int_{\mathbb{R}} |x|^2 \, \mu(\mathrm{d}x) < \infty$, a phase transition occurs for the critical parameter $p_c = 1/2$, in the sense that $\hat{S}$ is diffusive for $p > 1/2$ and superdiffusive for $p < 1/2$; see [10, 11]. In spite of the natural coupling between (1) and (2), there are thus major differences[3] between the asymptotic behaviors of $\check{S}$ and of $\hat{S}$: Proposition 1.1 shows that the asymptotic speed of $\check{S}$ depends on p, and Theorem 1.2 that there is no such phase transition for counterbalanced steps and $\check{S}$ is always diffusive.

The phase transition for step reinforcement when μ has a finite second moment can be explained informally as follows; for the sake of simplicity, suppose also that μ is centered, i.e. $m_1 = 0$. There are $\mathrm{i}(n) \sim pn$ trees in Simon's genealogical forest, which are overwhelmingly microscopic (i.e. of size $O(1)$), whereas only a few trees reach the size $O(n^{1-p})$. Because μ is centered, the contribution of microscopic trees to $\hat{S}(n)$ is of order $\sqrt{n}$, and that of the few largest trees of order n^{1-p}. This is the reason why $\hat{S}(n)$ grows like $\sqrt{n} \gg n^{1-p}$ when $p > 1/2$, and rather like $n^{1-p} \gg \sqrt{n}$ when $p < 1/2$. For counterbalanced steps, we will see that, due to the counterbalancing mechanism, the contribution of a large tree of size $\ell \gg 1$ is now only of order $\sqrt{\ell}$. As a consequence, the contribution to $\check{S}(n)$ of the largest trees of Simon's genealogical forest is only of order $O(n^{(1-p)/2})$. This is always much smaller than the contribution of microscopic trees which remain of order $\sqrt{n}$. We further stress that, even though only the sizes of the trees in Simon's genealogical forest are relevant for the analysis of the random walk $\hat{S}$ with reinforced steps, the study of the random walk $\check{S}$ with counterbalanced steps is more complex and requires information on the fine structure of those trees, not merely their sizes.

The rest of this text is organized as follows. Section 2 focusses on the purely counterbalanced case $p = 0$. In this situation, for each fixed $n \geq 1$, the distribution of $\check{S}(n)$ can be expressed explicitly in terms of Eulerian numbers. Section 3 is devoted to the coupling between the counterbalancing algorithm (1) and H. A. Simon's algorithm (2), and to the interpretation of the former in terms of a forest of random recursive trees induced by the latter. Proposition 1.1 and Theorem 1.2 are proved in Section 4, where we analyze more finely the respective contributions of some natural subfamilies. Last, in Section 5, we present a stable version of Theorem 1.2 when μ belongs to the domain of attraction (without centering) of an α-stable distribution for some $\alpha \in (0, 2)$.

[3] This should not come as a surprise. In the simplest case when $\mu = \delta_1$ is the Dirac mass at 1, one has $\hat{S}(n) \equiv n$, whereas $\check{S}$ is a truly stochastic process, even for $p = 0$ when there is no innovation.

2. Warm-up: the purely counterbalanced case

This section is devoted to the simpler situation[4] when $p = 0$. So $\varepsilon_n \equiv 0$ for all $n \geq 2$, meaning that every step, except of course the first one, counterbalances some preceding step. The law μ then only plays a superficial role as it is merely relevant for the first step. For the sake of simplicity, we further focus on the case when $\mu = \delta_1$ is the Dirac mass at 1.

The dynamics are entirely encoded by the sequence $(v(n))_{n\geq 2}$ of independent uniform variables on $\{1, \ldots, n-1\}$; more precisely the purely counterbalanced sequence of bits is given by

$$\check{X}_1 = 1 \quad \text{and} \quad \check{X}_n = -\check{X}_{v(n)} \quad \text{for all } n \geq 2. \tag{3}$$

The random algorithm (3) points to a convenient representation in terms of random recursive trees. Specifically, the sequence $(v(n))_{n\geq 2}$ encodes a random tree $\mathbb{T}_\infty$ with set of vertices $\mathbb{N}$ and set of edges $\{(v(n), n) : n \geq 2\}$. Roughly speaking, $\mathbb{T}_\infty$ is constructed recursively by incorporating vertices one after the other and creating an edge between each new vertex n and its parent $v(n)$ which is picked uniformly at random in $\{1, \ldots, n-1\}$ and independently of the other vertices. If we view 1 as the root of $\mathbb{T}_\infty$ and call a vertex j *odd* (respectively, *even*) when its generation (i.e. its distance to the root in $\mathbb{T}_\infty$) is an odd (respectively, even) number, then

$$\check{X}_n = \begin{cases} 1 & \text{if } n \text{ is an even vertex in } \mathbb{T}_\infty, \\ -1 & \text{if } n \text{ is an odd vertex in } \mathbb{T}_\infty. \end{cases}$$

Let us now introduce some relevant notation. For every $n \geq 1$, we write $\mathbb{T}_n$ for the restriction of $\mathbb{T}_\infty$ to the set of vertices $\{1, \ldots, n\}$ and refer to $\mathbb{T}_n$ as a *random recursive tree* of size n. We also write $\mathrm{Odd}(\mathbb{T}_n)$ (respectively, $\mathrm{Even}(\mathbb{T}_n)$) for the number of odd (respectively, even) vertices in $\mathbb{T}_n$ and set

$$\Delta(\mathbb{T}_n) := \mathrm{Even}(\mathbb{T}_n) - \mathrm{Odd}(\mathbb{T}_n) = n - 2\,\mathrm{Odd}(\mathbb{T}_n).$$

Of course, we can also express

$$\Delta(\mathbb{T}_n) = \check{X}_1 + \cdots + \check{X}_n,$$

which is the trajectory of an elephant full of regrets (i.e. for $q = 0$).

The main observation of this section is that the law of the number of odd vertices is readily expressed in terms of Eulerian numbers. Recall that $\left\langle {n \atop k} \right\rangle$ denotes the number of permutations ς of $\{1, \ldots, n\}$ with k descents, i.e. such that $\#\{1 \leq j < n : \varsigma(j) > \varsigma(j+1)\} = k$. Obviously $\left\langle {n \atop k} \right\rangle \geq 1$ if and only if $0 \leq k < n$, and one has

$$\sum_{k=0}^{n-1} \left\langle {n \atop k} \right\rangle = n!.$$

[4] Observe that this case without innovation has been excluded in Theorem 1.2.

The linear recurrence equation

$$\left\langle {n \atop k} \right\rangle = (n-k)\left\langle {n-1 \atop k-1} \right\rangle + (k+1)\left\langle {n-1 \atop k} \right\rangle \tag{4}$$

is easily derived from a recursive construction of permutations (see [30, Theorem 1.3]); we also mention the explicit formula (see [30, Corollary 1.3])

$$\left\langle {n \atop k} \right\rangle = \sum_{j=0}^{k} (-1)^j \binom{n+1}{j} (k+1-j)^n.$$

Lemma 2.1. *For every $n \geq 1$, we have*

$$\mathbb{P}(\mathrm{Odd}(\mathbb{T}_n) = \ell) = \frac{1}{(n-1)!} \left\langle {n-1 \atop \ell-1} \right\rangle,$$

with the convention[5] *that $\langle {0 \atop -1} \rangle = 1$ on the right-hand side for $n = 1$ and $\ell = 0$.*

Proof. Fix $n \geq 1$ and note that the very construction of random recursive trees gives the identity

$$\mathbb{P}(\mathrm{Odd}(\mathbb{T}_{n+1}) = \ell) = \frac{\ell}{n}\, \mathbb{P}(\mathrm{Odd}(\mathbb{T}_n) = \ell) + \frac{n+1-\ell}{n}\, \mathbb{P}(\mathrm{Odd}(\mathbb{T}_n) = \ell - 1).$$

Indeed, the first term of the sum on the right-hand side accounts for the event that the parent $v(n+1)$ of the new vertex $n+1$ is an odd vertex (then $n+1$ is an even vertex), and the second term for the event that $v(n+1)$ is an even vertex (then $n+1$ is an odd vertex).

In terms of $A(n,k) := n!\mathbb{P}(\mathrm{Odd}(\mathbb{T}_{n+1}) = k+1)$, this yields

$$A(n,k) = (k+1)A(n-1,k) + (n-k)A(n-1,k-1),$$

which is the linear recurrence equation (4) satisfied by the Eulerian numbers. Since plainly $A(1,0) = \mathbb{P}(\mathrm{Odd}(2) = 1) = 1 = \langle {1 \atop 0} \rangle$, we conclude by iteration that $A(n,k) = \langle {n \atop k} \rangle$ for all $n \geq 1$ and $0 \leq k < n$. Last, the formula in the statement also holds for $n = 1$ since $\mathrm{Odd}(1) = 0$. ∎

Remark 2.2. Lemma 2.1 is implicit in Mahmoud [24].[6] Indeed $\mathrm{Odd}(\mathbb{T}_n)$ can be viewed as the number of blue balls in an analytic Friedman's urn model started with one white ball and replacement scheme $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$; see [24, Section 7.2.2]. In this setting, Lemma 2.1 is equivalent to the formula for the number of white balls [24, bottom of p. 127]. Mahmoud relied on the analysis of the differential system associated to the replacement scheme via a Riccati differential equation and inversion of generating functions. The present approach

[5] Note that this convention is in agreement with the linear recurrence equation (4).

[6] Beware however that the definition of Eulerian numbers in [24] slightly differs from ours, namely $\langle {n \atop k} \rangle$ there corresponds to $\langle {n \atop k-1} \rangle$ here.

based on the linear recurrence equation (4) is more direct. Lemma 2.1 is also a closed relative to a result due to Najock and Heyde [27] (see also Mahmoud [24, Theorem 8.6] and Drmota [17, Section 6.2.4]) which states that the number of leaves in a random recursive tree of size n has the same distribution as that appearing in Lemma 2.1.

We next point to a useful identity related to Lemma 2.1 which goes back to Laplace (see Stanley [33, Chapter I, Exercise 51]) and is often attributed to Tanny [34]. For every $n \geq 0$, we have the identity in distribution

$$\mathrm{Odd}(\mathbb{T}_{n+1}) \overset{(\mathrm{d})}{=} \lceil U_1 + \cdots + U_n \rceil, \tag{5}$$

where on the right-hand side, $U_1, U_2, \ldots$ is a sequence of i.i.d. uniform variables on $[0, 1]$ and $\lceil \cdot \rceil$ denotes the ceiling function. We now record for future use the following consequences.

Corollary 2.3. (i) *For every $n \geq 2$, the random variable $\Delta(\mathbb{T}_n)$ is symmetric, that is, $\Delta(\mathbb{T}_n) \overset{(\mathrm{d})}{=} -\Delta(\mathbb{T}_n)$, and in particular $\mathbb{E}(\Delta(\mathbb{T}_n)) = 0$.*

(ii) *For all $n \geq 3$, one has $\mathbb{E}(\Delta(\mathbb{T}_n)^2) = n/3$.*

(iii) *For all $n \geq 1$, one has $\mathbb{E}(|\Delta(\mathbb{T}_n)|^4) \leq 6n^2$.*

Proof. (i) Equivalently, the assertion states that in a random recursive tree of size at least 2, the number of odd vertices and the number of even vertices have the same distribution. This is immediate from (5) and can also be checked directly from the construction.

(ii) This has already been observed by Schütz and Trimper [31] in the setting of the elephant random walk; for the sake of completeness we present a short argument. The vertex $n + 1$ is odd (respectively, even) in $\mathbb{T}_{n+1}$ if and only if its parent is an even (respectively, odd) vertex in $\mathbb{T}_n$. Hence one has

$$\mathbb{E}(\Delta(\mathbb{T}_{n+1}) - \Delta(\mathbb{T}_n) \mid \mathbb{T}_n) = -\frac{1}{n}\Delta(\mathbb{T}_n),$$

and since $\Delta(\mathbb{T}_{n+1}) - \Delta(\mathbb{T}_n) = \pm 1$, this yields the recursive equation

$$\mathbb{E}(\Delta(\mathbb{T}_{n+1})^2) = (1 - 2/n)\mathbb{E}(\Delta(\mathbb{T}_n)^2) + 1.$$

By iteration, we conclude that $\mathbb{E}(\Delta(\mathbb{T}_n)^2) = n/3$ for all $n \geq 3$.

(iii) Recall that the process of the fractional parts $\{U_1 + \cdots + U_n\}$ is a Markov chain on $[0, 1)$ whose distribution at any fixed time $n \geq 1$ is uniform on $[0, 1)$. Writing

$$V_n = 1 - 2U_n \quad \text{and} \quad W_n = 2\{U_1 + \cdots + U_n\} - 1,$$

we see that $V_1, V_2, \ldots$ is a sequence of i.i.d. uniform variables on $[-1, 1]$ and that W_n has the uniform distribution on $[-1, 1]$ too.

The characteristic function of the uniform variable V_j is

$$\mathbb{E}(\exp(\mathrm{i}\theta V_j)) = \theta^{-1}\sin(\theta) = 1 - \frac{\theta^2}{6} + \frac{\theta^4}{120} + O(\theta^6) \quad \text{as } \theta \to 0,$$

and therefore for every $n \geq 1$,

$$\mathbb{E}(\exp(\mathrm{i}\theta(V_1 + \cdots + V_n))) = \left(1 - \frac{\theta^2}{6} + \frac{\theta^4}{120} + O(\theta^6)\right)^n$$
$$= 1 - \frac{n}{6}\theta^2 + \left(\frac{n}{120} + \frac{n(n-1)}{72}\right)\theta^4 + O(\theta^6).$$

It follows that

$$\mathbb{E}((V_1 + \cdots + V_n)^4) = 24\left(\frac{n}{120} + \frac{n(n-1)}{72}\right) \leq n^2/3.$$

We can rephrase (5) as the identity in distribution

$$\Delta(\mathbb{T}_{n+1}) \overset{(\mathrm{d})}{=} V_1 + \cdots + V_n + W_n.$$

Since $\mathbb{E}(W_n^4) = 1/3$, the proof is completed with the elementary bound $(a+b)^4 \leq 16(a^4 + b^4)$. ∎

We now conclude this section with an application of (5) to the asymptotic normality of $\Delta(\mathbb{T}_n)$. Since $\mathbb{E}(U) = 1/2$ and $\mathrm{Var}(U) = 1/12$, the classical central limit theorem immediately yields the following.

Corollary 2.4. *Assume $p = 0$ and $\mu = \delta_1$. One has*

$$\lim_{n\to\infty} \frac{\Delta(\mathbb{T}_n)}{\sqrt{n}} = \mathcal{N}(0, 1/3) \quad \textit{in distribution.}$$

Corollary 2.4 goes back to [27] in the setting of the number of leaves in random recursive trees; see also [4,5,15,16] for alternative proofs in the framework of the elephant random walk.

3. Genealogical trees in Simon's algorithm

From now on, μ is an arbitrary probability law on $\mathbb{R}$ and we also suppose that the innovation rate is strictly positive, $p \in (0, 1)$. Recall the construction of the sequence $(\hat{X}_n)$ from Simon's reinforcement algorithm (2). Simon was interested in the asymptotic frequencies of variables having a given number of occurrences. Specifically, for every $n, j \in \mathbb{N}$, we write

$$N_j(n) := \#\{\ell \leq n : \hat{X}_\ell = X_j\},$$

for the number of occurrences of the variable X_j until the n-th step of the algorithm (2), and

$$\nu_k(n) := \#\{1 \leq j \leq \mathrm{i}(n) : N_j(n) = k\}, \quad k \in \mathbb{N}, \tag{6}$$

for the number of such variables that have occurred exactly k times. Observe also that the number of innovations satisfies the law of large numbers $\mathrm{i}(n) \sim pn$ a.s.

Lemma 3.1. *For every $k \geq 1$, we have*

$$\lim_{n\to\infty} \frac{\nu_k(n)}{pn} = \frac{1}{1-p}\mathrm{B}(k, 1 + 1/(1-p)) \quad \textit{in probability,}$$

where B *denotes the beta function.*

Lemma 3.1 is essentially due to H. A. Simon [32], who actually only established the convergence of the mean value. The strengthening to convergence in probability can be obtained as in [13] from a concentration argument based on Azuma–Hoeffding's inequality; see [28, Section 3.1]. The right-hand side in the formula is a probability mass on $\mathbb{N}$ known as the *Yule–Simon distribution* with parameter $1/(1-p)$. We record for future use a couple of identities which are easily checked from the integral definition of the beta function:

$$\frac{1}{1-p}\sum_{k=1}^{\infty} \mathrm{B}(k, 1 + 1/(1-p)) = 1 \tag{7}$$

and

$$\frac{1}{1-p}\sum_{k=1}^{\infty} k\mathrm{B}(k, 1 + 1/(1-p)) = \frac{1}{p}. \tag{8}$$

For $k = 1$, Lemma 3.1 reads

$$\lim_{n\to\infty} n^{-1}\nu_1(n) = \frac{p}{2-p} \quad \text{in probability.} \tag{9}$$

We shall also need to estimate the fluctuations, which can be derived by specializing a Gaussian limit theorem for extended Pólya urns due to Bai et al. [1].

Lemma 3.2. *We have*

$$\lim_{n\to\infty} \frac{\nu_1(n) - np/(2-p)}{\sqrt{n}} = \mathcal{N}\left(0, \frac{2p^3 - 8p^2 + 6p}{(3-2p)(2-p)^2}\right) \quad \textit{in distribution.}$$

Proof. The proof relies on the observation that Simon's algorithm can be coupled with a two-color urn governed by the same sequences (ε_n) of random bits and $(v(n))$ of uniform variables as follows. Imagine that we observe the outcome of Simon's algorithm at the n-th step and that for each $1 \leq j \leq n$, we associate a white ball if the variable $\hat{X}_j$ appears exactly once, and a red ball otherwise. At the initial time $n = 1$, the urn contains just one white ball and no red balls. At each step $n \geq 2$, a ball picked uniformly at random in the urn (in terms of Simon's algorithm, this is given by the uniform variable $v(n)$). If $\varepsilon_n = 1$, then the ball picked is returned to the urn and one adds a white ball (in terms of Simon's algorithm, this corresponds to an innovation and $\nu_1(n) = \nu_1(n-1) + 1$). If $\varepsilon_n = 0$, then the ball picked is removed from the urn and one adds two red balls (in terms of Simon's algorithm, this corresponds to a repetition and either $\nu_1(n) = \nu_1(n-1) - 1$ if the ball picked is white, or $\nu_1(n) = \nu_1(n-1)$ if the ball picked is red). By construction, the number W_n of white balls in the urn coincides with the number $\nu_1(n)$ of variables that have appeared exactly once in Simon's algorithm (2).

We shall now check our claim in the setting of [1] by specifying the quantities which appear there. The evolution of the number of white balls in the urn is governed by equation (2.1) in [1], viz.

$$W_n = W_{n-1} + I_n A_n + (1 - I_n)C_n,$$

where $I_n = 1$ if a white ball is picked and $I_n = 0$ otherwise. In our framework, we further have $A_n = 2\varepsilon_n - 1$ and $C_n = \varepsilon_n$. If we write $\mathcal{F}_n$ for the natural filtration generated by the variables $(A_k, C_k, I_k)_{k \leq n}$, then A_n and C_n are independent of $\mathcal{F}_{n-1}$ with

$$\mathbb{E}(A_n) = 2p - 1, \quad \mathbb{E}(C_n) = p, \quad \mathrm{Var}(A_n) = 4(p - p^2), \quad \mathrm{Var}(C_n) = p - p^2.$$

This gives, in the notation of [1, (2.2)],

$$\begin{aligned}\sigma_M^2 &= \frac{p}{2-p} 4(p - p^2) + \left(1 - \frac{p}{2-p}\right)(p - p^2) + (p-1)^2 \frac{p}{2-p}\left(1 - \frac{p}{2-p}\right) \\ &= \frac{2p^3 - 8p^2 + 6p}{(2-p)^2},\end{aligned}$$

and finally

$$\sigma^2 = \frac{2p^3 - 8p^2 + 6p}{(3-2p)(2-p)^2}.$$

Our claim can now be seen as a special case of [1, Corollary 2.1]. ∎

We shall also need a refinement of Lemma 3.1 in which one records not only the number of occurrences of the variable X_j, but more generally the genealogical structure of these occurrences. We need to introduce some notation first.

Fix $n \geq 1$ and $1 \leq j \leq \mathrm{i}(n)$ (i.e. the variable X_j has already appeared at the n-th step of the algorithm). Write $\ell_1 < \cdots < \ell_k \leq n$ for the increasing sequence of steps of the algorithm at which X_j appears, where $k = N_j(n) \geq 1$. The genealogy of occurrences of the variable X_j until the n-th step is recorded as a tree $T_j(n)$ on $\{1, \ldots, k\}$ such that for every $1 \leq a < b \leq k$, (a, b) is an edge of $T_j(n)$ if and only if $v(\ell_b) = \ell_a$, that is, if and only if the identity $\hat{X}_{\ell_b} = X_j$ actually results from the fact that the algorithm repeats the variable $\hat{X}_{\ell_a}$ at its ℓ_b-th step. Plainly, $T_j(n)$ is an increasing tree of size k, meaning a tree on $\{1, \ldots, k\}$ such that the sequence of vertices along any branch from the root 1 to a leaf is increasing. In this direction, we recall that there are $(k-1)!$ increasing trees of size k and that the uniform distribution of the set of increasing trees of size k coincides with the law of $\mathbb{T}_k$ the random recursive tree of size k. See for instance Drmota [17, Section 1.3.1].

More generally, the distribution of the entire genealogical forest given the sizes of the genealogical trees can be described as follows.

Lemma 3.3. *Fix $n \geq 1$, $1 \leq k \leq n$, and let $n_1, \ldots, n_k \geq 1$ with $n_1 + \cdots + n_k = n$. Then conditionally on $N_j(n) = n_j$ for every $j = 1, \ldots, k$, the genealogical trees $T_1(n), \ldots, T_k(n)$ are independent random recursive trees of respective sizes $n_1, \ldots, n_k$.*

Proof. Recall that $\{(v(j), j)\colon 1 \leq j \leq n\}$ is the set of edges of $\mathbb{T}_n$, the random recursive tree of size n. The well-known splitting property states that removing a given edge

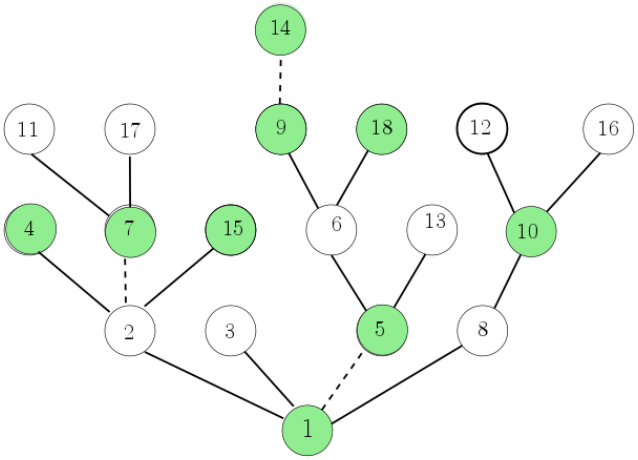

Fig. 1. Example of a genealogical forest representation of Simon's algorithm (2) after 18 steps. The dotted edges account for innovation events, i.e. $\varepsilon_j = 1$, and the four genealogical trees are rooted at $1, 5, 7, 14$. In each subtree, vertices at even generations are colored green and vertices at odd generations are white. For instance the genealogical tree $T_2(18)$ is rooted at 5, it has three even vertices and two odd vertices.

from $\mathbb{T}_n$, say $(v(j), j)$ for some fixed j, produces two subtrees, which in turn, conditionally on their sizes, are two independent random recursive trees. This has been observed first by Meir and Moon [25]; see also [3] and references therein for more about this property.

The genealogical trees $T_1(n), \ldots, T_k(n)$ result by removing from $\mathbb{T}_n$ the edges $(v(j), j)$ for which $\varepsilon_j = 1$ and enumerating in each subtree component their vertices in increasing order. Our statement is now easily seen by applying this splitting property iteratively. ∎

For the proofs of Proposition 1.1 and Theorem 1.2 we shall also need an argument of uniform integrability, which relies in turn on the following lemma. Recall that if T is a rooted tree, $\Delta(T)$ denotes the difference between the number of vertices at even distance from the root and that at odd distance.

Lemma 3.4. *For every* $1 < \beta < 2 \wedge \frac{1}{1-p}$, *one has*

$$\sup_{n\geq 1} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}(N_j(n)^\beta) < \infty$$

and

$$\sup_{n\geq 1} \frac{1}{n} \mathbb{E}\Big(\sum_{j=1}^{\mathrm{i}(n)} |\Delta(T_j(n))|^{2\beta}\Big) < \infty.$$

Proof. The first claim is a consequence of [9, Lemma 3.6] which states that for β in $(1, 1/(1-p))$ [beware that the parameter denoted by p in [9] is actually $1-p$ here], there exist numerical constants $c > 0$ and $\eta \in (0,1)$ such that $\mathbb{E}(N_j(n)^\beta) \le c(n/j)^\eta$ for all $1 \le j \le n$.

Next, combining Jensen's inequality with Corollary 2.3 (iii), we find that for $k \ge 2$,

$$\mathbb{E}(|\Delta(\mathbb{T}_k)|^{2\beta}) \le \mathbb{E}(|\Delta(\mathbb{T}_k)|^4)^{\beta/2} \le 6k^\beta.$$

Then recall that conditionally on $N_j(n) = k \ge 1$, $T_j(n)$ has the law of $\mathbb{T}_k$, the random recursive tree of size k, and hence

$$\begin{aligned}\mathbb{E}\Big(\sum_{j=1}^{\mathrm{i}(n)} |\Delta(T_j(n))|^{2\beta}\Big) &= \sum_{j=1}^{n}\Big(\sum_{k=1}^{n} \mathbb{E}(|\Delta(\mathbb{T}_k)|^{2\beta})\mathbb{P}(N_j(n)=k)\Big)\\ &\le 6\sum_{j=1}^{n}\Big(\sum_{k=1}^{n} k^\beta\,\mathbb{P}(N_j(n)=k)\Big).\end{aligned}$$

We know from the first part that this last quantity is finite, and the proof is complete. ∎

4. Proofs of the main results

As its title indicates, the purpose of this section is to establish Proposition 1.1 and Theorem 1.2. The observation that for every $n \ge 1$ and $1 \le j \le \mathrm{i}(n)$, the variable X_j appears exactly $\mathrm{Even}(T_j(n))$ times and its opposite $-X_j$ exactly $\mathrm{Odd}(T_j(n))$ times until the n-th step of the algorithm (1), yields the identity

$$\check{S}(n) := \sum_{i=1}^{n} \check{X}_i = \sum_{j=1}^{\mathrm{i}(n)} \Delta(T_j(n))X_j, \tag{10}$$

which lies at the heart of our approach. We stress that in (10) as well as in related expressions that we shall use below, the sequence (X_n) of i.i.d. variables and the family $(T_j(n))$ of genealogical trees are independent, because the latter are constructed from the sequences (ε_n) and $(v(n))$ only.

Actually, our proof analyzes more precisely the effects of the counterbalancing algorithm (1) by estimating specifically the contributions of certain subfamilies to the asymptotic behavior of $\check{S}$. Specifically, we set, for every $k \ge 1$,

$$\check{S}_k(n) := \sum_{j=1}^{\mathrm{i}(n)} \Delta(T_j(n))X_j \mathbb{1}_{N_j(n)=k}, \tag{11}$$

so that

$$\check{S}(n) = \sum_{k=1}^{n} \check{S}_k(n).$$

4.1. Proof of Proposition 1.1

The case $p = 1$ (no counterbalancing events) of Proposition 1.1 is just the weak law of large numbers, and the case $p = 0$ (no innovations) is a consequence of Corollary 2.4. The case $p \in (0, 1)$ derives from the next lemma which shows more precisely that the variables X_j that have appeared in the algorithm (1) but have not yet counterbalanced determine the ballistic behavior of $\check{S}$, whereas those that have appeared twice or more (i.e. such that $N_j(n) \geq 2$) have a negligible impact.

Lemma 4.1. *Assume that $X_1 \in L^1(\mathbb{P})$ and recall that $m_1 = \mathbb{E}(X_1)$. Then the following limits hold in probability:*

(i) $\lim_{n\to\infty} n^{-1}\check{S}_1(n) = m_1 p/(2-p)$,

(ii) $\lim_{n\to\infty} n^{-1}\sum_{k=2}^{n} |\check{S}_k(n)| = 0$.

Proof. (i) Recall the notation (6) and that the sequence (X_j) of i.i.d. variables is independent of the events $\{N_j(n) = 1\}$. We have the identity in distribution

$$\check{S}_1(n) \overset{(d)}{=} S_1(\nu_1(n)),$$

where $S_1(n) = X_1 + \cdots + X_n$ is the usual random walk. Claim (i) follows readily from the law of large numbers and (9).

(ii) We first argue that for each fixed $k \geq 2$,

$$\lim_{n\to\infty} n^{-1}\check{S}_k(n) = 0 \quad \text{almost surely.} \tag{12}$$

Indeed, recall that $\nu_k(n)$ denotes the number of genealogical trees $T_j(n)$ of size k. It follows from Lemma 3.3 that conditionally on $\nu_k(n) = \ell$, the subfamily of such $T_j(n)$, enumerated in the increasing order of the index j, is given by ℓ i.i.d. copies of the random recursive tree $\mathbb{T}_k$. Hence, still conditionally on $\nu_k(n) = \ell$, enumerating the elements of the subfamily $\{X_j\Delta(T_j(n)) : N_j(n) = k\}$ in the increasing order of j yields ℓ independent variables, each being distributed as $X_1\Delta(\mathbb{T}_k)$ with X_1 and $\Delta(\mathbb{T}_k)$ independent. We deduce from Corollary 2.3 (i) that the variable $X_1\Delta(\mathbb{T}_k)$ is symmetric, and since it is also integrable, it is centered. Since $\nu_k(n) \leq n$, this readily entails (12) by an application of the law of large numbers.

The proof can be completed by an argument of uniform integrability. In this direction, fix an arbitrarily large integer ℓ and write, by the triangular inequality,

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\Big(\sum_{k=\ell}^{n} |\check{S}_k(n)|\Big) &\leq \frac{1}{n}\mathbb{E}\Big(\sum_{j=1}^{\mathrm{i}(n)} |X_j| N_j(n)\mathbb{1}_{N_j(n)\geq\ell}\Big) \\
&= \frac{\mathbb{E}(|X_1|)}{n}\sum_{j=1}^{n}\mathbb{E}(N_j(n)\mathbb{1}_{N_j(n)\geq\ell}) \\
&\leq \ell^{1-\beta}\frac{\mathbb{E}(|X_1|)}{n}\sum_{j=1}^{n}\mathbb{E}(N_j(n)^{\beta}),
\end{aligned}$$

where the last inequality holds for any $\beta > 1$. We see from Lemma 3.4 that the right-hand side converges to 0 as $\ell \to \infty$ uniformly in $n \geq 1$, and the rest of the proof is straightforward. ∎

4.2. Proof of Theorem 1.2

For $p = 1$ (no counterbalancing events), Theorem 1.2 just reduces to the classical central limit theorem, so we assume $p \in (0, 1)$. The first step of the proof consists in determining jointly the fluctuations of the components $\check{S}_k$ defined in (11).

Lemma 4.2. *Assume that $m_2 = \mathbb{E}(X_1^2) < \infty$. Then as $n \to \infty$, the sequences of random variables*

$$\frac{\check{S}_1(n) - pm_1/(2-p)}{\sqrt{n}} \quad (\textit{for } k = 1)$$

and

$$\frac{\check{S}_k(n)}{\sqrt{n}} \quad (\textit{for } k \geq 2)$$

converge jointly in distribution towards a sequence

$$(\mathcal{N}_k(0, \sigma_k^2))_{k \geq 1}$$

of independent centered Gaussian variables, where

$$\sigma_1^2 := \frac{pm_2}{2-p} - \frac{p^2 m_1^2}{(3-2p)(2-p)^2},$$

$\sigma_2^2 := 0$, *and*

$$\sigma_k^2 := \frac{kpm_2}{3(1-p)} \mathrm{B}(k, 1 + 1/(1-p)) \quad \textit{for } k \geq 3.$$

Proof. For each $k \geq 1$, let $(Y_k(n))_{n \geq 1}$ be a sequence of i.i.d. copies of $\Delta(\mathbb{T}_k)X$, where X has the law μ and is independent of the random recursive tree $\mathbb{T}_k$. We further assume that these sequences are independent. Taking partial sums yields a sequence indexed by k of independent random walks

$$S_k(n) = Y_k(1) + \cdots + Y_k(n), \quad n \geq 0.$$

For each $n \geq 1$, the family of blocks

$$B_k(n) := \{j \leq \mathrm{i}(n) : N_j(n) = k\} \quad \text{for } 1 \leq k \leq \mathrm{i}(n)$$

forms a random partition of $\{1, \ldots, \mathrm{i}(n)\}$ which is independent of the X_j's. Recall that we are using the notation $\nu_k(n) = \#B_k(n)$, and also from Lemma 3.3, that conditionally on the $N_j(n)$'s, the genealogical trees $T_j(n)$ are independent random recursive trees. We now deduce from the very definition (11), for every fixed $n \geq 1$, the identity in distribution

$$(\check{S}_k(n))_{k \geq 1} \overset{(\mathrm{d})}{=} (S_k(\nu_k(n)))_{k \geq 1},$$

where on the right-hand side, the random walks $(S_k)_{k \geq 1}$ are independent of the sequence $(\nu_k(n))_{k \geq 1}$ of block sizes.

Next we write, first for $k = 1$,

$$S_1(\nu_1(n)) - \frac{pn}{2-p}m_1 = S_1\left(\left\lfloor \frac{pn}{2-p} \right\rfloor\right) - \frac{pn}{2-p}m_1 + \sum_{j=\lceil pn/(2-p)\rceil}^{\nu_1(n)} Y_1(j),$$

second $S_2 \equiv 0$ (since $\Delta(\mathbb{T}_2) \equiv 0$) for $k = 2$, and third, for $k \geq 3$,

$$S_k(\nu_k(n)) = S_k\left(\left\lfloor \frac{pn}{1-p}\mathrm{B}(k, 1 + 1/(1-p)) \right\rfloor\right) + \sum_{j=\lceil \frac{pn}{1-p}\mathrm{B}(k,1+1/(1-p))\rceil}^{\nu_k(n)} Y_k(j),$$

with the usual convention that $\sum_{j=a}^{b} = -\sum_{j=b}^{a}$ when $b < a$.

Since the i.i.d. variables $Y_1(\cdot)$ have mean m_1 and variance $m_2 - m_1^2$, the central limit theorem ensures the convergence in distribution

$$\lim_{n\to\infty} n^{-1/2}\left(S_1\left(\left\lfloor \frac{pn}{2-p} \right\rfloor\right) - \frac{pn}{2-p}m_1\right) = \mathcal{N}_1\left(0, \frac{p(m_2 - m_1^2)}{2-p}\right). \tag{13}$$

Similarly, for $k \geq 3$, each $Y_k(n)$ is centered with variance $km_2/3$ (by Corollary 2.3 (i, ii)) and hence, using the notation in the statement, we get the convergence in distribution

$$\lim_{n\to\infty} n^{-1/2} S_k\left(\left\lfloor \frac{pn}{1-p}\mathrm{B}(k, 1 + 1/(1-p)) \right\rfloor\right) = \mathcal{N}_k(0, \sigma_k^2). \tag{14}$$

Plainly, the weak convergences (13) and (14) hold jointly when we agree that the limits are independent Gaussian variables.

Next, from Lemma 3.1 and the fact that for $k \geq 3$, the i.i.d. variables $Y_k(j)$ are centered with finite variance, we easily get

$$\lim_{n\to\infty} n^{-1/2}\left| \sum_{j=\lceil \frac{pn}{1-p}\mathrm{B}(k,1+1/(1-p))\rceil}^{\nu_k(n)} Y_k(j)\right| = 0 \quad \text{in } L^2(\mathbb{P}).$$

Finally, for $k = 1$, we write

$$\sum_{j=\lceil pn/(2-p)\rceil}^{\nu_1(n)} Y_1(j) = m_1(\nu_1(n) - \lfloor pn/(2-p)\rfloor) + \sum_{j=\lceil pn/(2-p)\rceil}^{\nu_1(n)} (Y_1(j) - m_1).$$

On the one hand, from the same argument as above we have

$$\lim_{n\to\infty} n^{-1/2}\left| \sum_{j=\lceil pn/(2-p)\rceil}^{\nu_1(n)} (Y_1(j) - m_1)\right| = 0 \quad \text{in } L^2(\mathbb{P}).$$

On the other hand, Lemma 3.2 already implies the convergence in distribution

$$\lim_{n\to\infty} m_1 \frac{\nu_1(n) - \lfloor pn/(2-p)\rfloor}{\sqrt{n}} = \mathcal{N}\left(0, \frac{2p^3 - 8p^2 + 6p}{(3-2p)(2-p)^2}m_1^2\right).$$

Obviously, this convergence in law holds jointly with (13) and (14), where the limiting Gaussian variables are independent. Putting the pieces together completes the proof. ∎

The final step for the proof of Theorem 1.2 is the following lemma.

Lemma 4.3. *We have*

$$\lim_{K\to\infty} \sup_{n\geq 1} n^{-1}\mathbb{E}\Big(\Big|\sum_{k\geq K} \check{S}_k(n)\Big|^2\Big) = 0.$$

Proof. We write

$$\sum_{k\geq K} \check{S}_k(n) = \sum_{j=1}^{n} X_j \Delta(T_j(n))\mathbb{1}_{N_j(n)\geq K}.$$

Since the X_j are independent of the $T_j(n)$, we get

$$\begin{aligned}\mathbb{E}\Big(\Big|\sum_{k\geq K} \check{S}_k(n)\Big|^2\Big) &= \mathbb{E}\Big(\sum_{j,j'=1}^{n} X_j X_{j'} \Delta(T_j(n))\mathbb{1}_{N_j(n)\geq K}\Delta(T_{j'}(n))\mathbb{1}_{N_{j'}(n)\geq K}\Big)\\ &\leq m_2 \sum_{j,j'=1}^{n} \mathbb{E}\big(\Delta(T_j(n))\mathbb{1}_{N_j(n)\geq K}\Delta(T_{j'}(n))\mathbb{1}_{N_{j'}(n)\geq K}\big).\end{aligned}$$

We evaluate the expectation on the right-hand side by conditioning first on $N_j(n) = k$ and $N_{j'}(n) = k'$ with $k, k' \geq 3$. Recall from Lemma 3.3 that the genealogical trees $T_j(n)$ and $T_{j'}(n)$ are then two random recursive trees with respective sizes k and k', which are further independent when $j \neq j'$. Thanks to Corollary 2.3 (i, ii) we get

$$\mathbb{E}\big(\Delta(T_j(n))\mathbb{1}_{N_{j'}(n)\geq K}\Delta(T_{j'}(n))\mathbb{1}_{N_{j'}(n)\geq K}\big) = \begin{cases} \frac{1}{3}\mathbb{E}(N_j(n)\mathbb{1}_{N_j(n)\geq K}) & \text{if } j = j',\\ 0 & \text{if } j \neq j'.\end{cases}$$

We have thus shown that

$$\mathbb{E}\Big(\Big|\sum_{k\geq K} \check{S}_k(n)\Big|^2\Big) \leq \frac{m_2}{3}\sum_{j=1}^{n} \mathbb{E}(N_j(n)\mathbb{1}_{N_j(n)\geq K}),$$

which yields our claim just as in the proof of Lemma 4.1 (ii). ∎

The proof of Theorem 1.2 is now easily completed by combining Lemmas 4.2 and 4.3. Indeed, the identity

$$\frac{pm_2}{2-p} + \sum_{k=2}^{\infty} \sigma_k^2 = \frac{m_2}{3-2p}$$

is easily checked from (8).

5. A stable central limit theorem

The arguments for the proof of Theorem 1.2 when the step distribution μ has a finite second moment can be adapted to the case when μ belongs to some stable domain of attraction; for the sake of simplicity we focus on the situation without centering. Specif-

ically, let (a_n) be a sequence of positive real numbers that is regularly varying with exponent $1/\alpha$ for some $\alpha \in (0,2)$, in the sense that $\lim_{n\to\infty} a_{\lfloor rn\rfloor}/a_n = r^{1/\alpha}$ for every $r > 0$, and suppose that

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{a_n} = Z \quad \text{in distribution,} \tag{15}$$

where Z is some α-stable random variable. We refer to [19, Theorems 4 and 5, pp. 181–182] and of [21, Chapter 2, Section 6] for necessary and sufficient conditions for (15) in terms of μ only. We write φ_α for the characteristic exponent of Z, viz.

$$\mathbb{E}(\exp(\mathrm{i}\theta Z)) = \exp(-\varphi_\alpha(\theta)) \quad \text{for all } \theta \in \mathbb{R};$$

recall that φ_α is homogeneous with exponent α, i.e.

$$\varphi_\alpha(\theta) = |\theta|^\alpha \varphi_\alpha(\mathrm{sgn}(\theta)) \quad \text{for all } \theta \neq 0.$$

Recall the definition and properties of the Eulerian numbers $\langle {n \atop k} \rangle$ from Section 2, and also the Pochhammer notation

$$(x)^{(k)} := \frac{\Gamma(x+k)}{\Gamma(x)} = \prod_{j=0}^{k-1} (x+j), \quad x > 0,\ k \in \mathbb{N},$$

for the rising factorial, where Γ stands for the gamma function. We can now claim:

Theorem 5.1. *Assume* (15). *For each* $p \in (0,1)$, *we have*

$$\lim_{n\to\infty} \frac{\check{S}(n)}{a_n} = \check{Z} \quad \textit{in distribution,}$$

where $\check{Z}$ *is an* α*-stable random variable with characteristic exponent* $\check{\varphi}_\alpha$ *given by*

$$\check{\varphi}_\alpha(\theta) = \frac{p}{1-p} \sum_{k=1}^{\infty} \sum_{\ell=0}^{k-1} \frac{\varphi_\alpha((k-2\ell)\theta)}{(1+1/(1-p))^{(k)}} \left\langle {k-1 \atop \ell-1} \right\rangle, \quad \theta \in \mathbb{R}.$$

The proof of Theorem 5.1 relies on a refinement of Simon's result (Lemma 3.1) to the asymptotic frequencies of genealogical trees induced by the reinforcement algorithm (2). We denote by $\mathcal{T}^\uparrow$ the set of increasing trees (of arbitrary finite size), and for any $\tau \in \mathcal{T}^\uparrow$, we write $|\tau|$ for its size (the number of vertices) and $\Delta(\tau)$ for the difference between its numbers of even vertices and of odd vertices. Refining (6), we also define

$$\nu_\tau(n) := \sum_{j=1}^{\mathrm{i}(n)} \mathbb{1}_{T_j(n)=\tau}, \quad \tau \in \mathcal{T}^\uparrow.$$

Lemma 5.2. *We have the identity*

$$\sum_{\tau\in\mathcal{T}^\uparrow} \frac{|\tau| + |\Delta(\tau)|^2}{(1+1/(1-p))^{(|\tau|)}} = \frac{4p}{3(1-p)},$$

and the convergence in probability

$$\lim_{n\to\infty} \sum_{\tau\in\mathcal{T}^{\uparrow}} (|\tau| + |\Delta(\tau)|^2) \left| \frac{\nu_\tau(n)}{n} - \frac{p}{(1-p)(1+1/(1-p))^{(|\tau|)}} \right| = 0.$$

Proof. We start by showing that for every $k \geq 1$, and every tree $\tau \in \mathcal{T}^{\uparrow}$ of size $|\tau| = k$, we have

$$\lim_{n\to\infty} \frac{\nu_\tau(n)}{n} = \frac{p}{(1-p)(1+1/(1-p))^{(k)}} \quad \text{in probability.} \tag{16}$$

Indeed, the distribution of the random recursive tree $\mathbb{T}_k$ of size k is the uniform probability measure on the set of increasing trees of size k, which has $(k-1)!$ elements. We deduce from Lemma 3.3 and the law of large numbers that

$$\nu_\tau(n) \sim \nu_k(n)/(k-1)!.$$

The claim (16) now follows from Lemma 3.1 and the identity

$$\mathrm{B}(k, 1+1/(1-p)) = \frac{(k-1)!}{(1+1/(1-p))^{(k)}}.$$

We now have to prove that (16) holds in $L^1(|\tau| + |\Delta(\tau)|^2, \mathcal{T}^{\uparrow})$. On the one hand, for every $n \geq 1$ one obviously has

$$\sum_{\tau\in\mathcal{T}^{\uparrow}} |\tau| \nu_\tau(n) = n.$$

On the other hand, there are $(k-1)!$ increasing trees of size k and hence

$$\frac{p}{1-p} \sum_{\tau\in\mathcal{T}^{\uparrow}} \frac{|\tau|}{(1+1/(1-p))^{(|\tau|)}} = \frac{p}{1-p} \sum_{k=1}^{\infty} k\mathrm{B}(k, 1+1/(1-p)) = 1,$$

where the second equality is (8). We deduce from Scheffé's Lemma and (16) the convergence in probability

$$\lim_{n\to\infty} \sum_{\tau\in\mathcal{T}^{\uparrow}} |\tau| \left| \frac{\nu_\tau(n)}{n} - \frac{p}{(1-p)(1+1/(1-p))^{(|\tau|)}} \right| = 0.$$

Similarly, we deduce from Corollary 2.3 (ii) and Lemma 3.3 that, for every $n \geq 0$,

$$\mathbb{E}\Big(\sum_{\tau\in\mathcal{T}^{\uparrow}} \Delta(\tau)^2 \nu_\tau(n) \Big) = \mathbb{E}\Big(\sum_{j=1}^{\mathrm{i}(n)} \Delta(T_j(n))^2 \Big) = \frac{1}{3}\mathbb{E}\Big(\sum_{j=1}^{\mathrm{i}(n)} |T_j(n)| \Big) = n/3,$$

and further, since there are $(k-1)!$ increasing trees of size k and $\mathbb{T}_k$ has the uniform distribution on the set of such trees,

$$\begin{aligned} \frac{p}{1-p} \sum_{\tau\in\mathcal{T}^{\uparrow}} \frac{\Delta(\tau)^2}{(1+1/(1-p))^{(|\tau|)}} &= \frac{p}{1-p} \sum_{k=1}^{\infty} \mathbb{E}(\Delta(\mathbb{T}_k)^2)\mathrm{B}(k, 1+1/(1-p)) \\ &= \frac{p}{1-p} \sum_{k=1}^{\infty} \frac{k}{3}\mathrm{B}(k, 1+1/(1-p)) = \frac{1}{3}. \end{aligned}$$

We conclude again from Scheffé's Lemma that

$$\lim_{n\to\infty} \sum_{\tau\in\mathcal{T}^\uparrow} \Delta(\tau)^2 \left| \frac{\nu_\tau(n)}{pn} - \frac{p}{(1-p)(1+1/(1-p))^{(|\tau|)}} \right| = 0,$$

and the proof is complete. ∎

We now establish Theorem 5.1.

Proof of Theorem 5.1. We denote the characteristic function of μ by

$$\Phi(\theta) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\theta x}\, \mu(\mathrm{d}x) \quad \text{for } \theta \in \mathbb{R}.$$

Fix $r > 0$ small enough so that $|1 - \Phi(\theta)| < 1$ whenever $|\theta| \le r$, and then define the characteristic exponent $\varphi : [-r, r] \to \mathbb{C}$ as the continuous determination of the logarithm of Φ on $[-r, r]$. In words, φ is the unique continuous function on $[-r, r]$ with $\varphi(0) = 0$ and such that $\Phi(\theta) = \exp(-\varphi(\theta))$ for all $\theta \in [-r, r]$. For definiteness, we further set $\varphi(\theta) = 0$ whenever $|\theta| > r$.

Next, observe from Markov's inequality that for any $1 < \beta < 2 \wedge (1-p)^{-1}$ and any $a > 0$,

$$\mathbb{P}\big(\exists j \le \mathrm{i}(n) : |\Delta(T_j(n))| \ge a\sqrt{n}\big) \le a^{-2\beta} n^{-\beta}\, \mathbb{E}\Big(\sum_{j=1}^{\mathrm{i}(n)} |\Delta(T_j(n))|^{2\beta}\Big),$$

so that, thanks to Lemma 3.4,

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \max_{1\le j\le \mathrm{i}(n)} |\Delta(T_j(n))| = 0 \quad \text{in probability.}$$

In particular, since the sequence (a_n) is regularly varying with exponent $1/\alpha > 1/2$, for every $\theta \in \mathbb{R}$ the events

$$\Lambda(n, \theta) := \{|\theta \Delta(T_j(n))/a_n| < r \text{ for all } j = 1, \ldots, \mathrm{i}(n)\}, \quad n \ge 1,$$

occur with high probability as $n \to \infty$, in the sense that $\lim_{n\to\infty} \mathbb{P}(\Lambda(n, \theta)) = 1$.

We then deduce from (10) and the fact that the variables X_j are i.i.d. with law μ that for every $\theta \in \mathbb{R}$,

$$\mathbb{E}\big(\exp(\mathrm{i}\theta \check{S}(n)/a_n) \mathbb{1}_{\Lambda(n,\theta)}\big) = \mathbb{E}\left(\exp\left(-\frac{1}{n}\sum_{j=1}^{\mathrm{i}(n)} n\varphi\big(\theta a_n^{-1} \Delta(T_j(n))\big)\right) \mathbb{1}_{\Lambda(n,\theta)}\right).$$

We then write, in the notation of Lemma 5.2,

$$\frac{1}{n}\sum_{j=1}^{\mathrm{i}(n)} n\varphi\big(\theta a_n^{-1} \Delta(T_j(n))\big) = \sum_{\tau\in\mathcal{T}^\uparrow} n\varphi\big(\theta a_n^{-1} \Delta(\tau)\big) \frac{\nu_\tau(n)}{n}.$$

Recall that we are assuming (15). According to Theorem 2.6.5 of Ibragimov and Linnik [21], φ is regularly varying at 0 with exponent α, and since $\alpha < 2$, the Potter bounds (see [12, Theorem 1.5.6]) show that for some constant C,

$$n|\varphi(\theta a_n^{-1}\Delta(\tau))| \le C|\theta\Delta(\tau)|^2. \tag{17}$$

We deduce from Lemma 5.2, for every fixed $\theta \in \mathbb{R}$, the convergence in probability

$$\lim_{n\to\infty} \sum_{\tau\in\mathcal{T}^\uparrow} n|\varphi(\theta a_n^{-1}\Delta(\tau))| \left| \frac{\nu_\tau(n)}{n} - \frac{p}{(1-p)(1+1/(1-p))^{(|\tau|)}} \right| = 0.$$

Furthermore, still from [21, Theorem 2.6.5], we have

$$\lim_{n\to\infty} n\varphi(\theta/a_n) = \varphi_\alpha(\theta) \quad \text{for every } \theta \in \mathbb{R},$$

and we deduce by dominated convergence, using Lemma 5.2 and (17), that

$$\lim_{n\to\infty} \sum_{\tau\in\mathcal{T}^\uparrow} |n\varphi(\theta a_n^{-1}\Delta(\tau)) - \varphi_\alpha(\theta\Delta(\tau))| \frac{p}{(1-p)(1+1/(1-p))^{(|\tau|)}} = 0.$$

Putting the pieces together, we have shown that

$$\lim_{n\to\infty} \mathbb{E}(\exp(\mathrm{i}\theta\check{S}(n)/a_n)) = \exp\left(- \sum_{\tau\in\mathcal{T}^\uparrow} \varphi_\alpha(\theta\Delta(\tau)) \frac{p}{(1-p)(1+1/(1-p))^{(|\tau|)}} \right).$$

It only remains to check that the right-hand side above agrees with the formula of the statement. This follows from Lemma 2.1 and the fact that for every $k \ge 1$, $\mathbb{T}_k$ has the uniform distribution on $\{\tau \in \mathcal{T}^\uparrow : |\tau| = k\}$. ∎

References

[1] Bai, Z. D., Hu, F., Zhang, L.-X.: Gaussian approximation theorems for urn models and their applications. Ann. Appl. Probab. **12**, 1149–1173 (2002) Zbl 1014.60025 MR 1936587

[2] Baur, E.: On a class of random walks with reinforced memory. J. Statist. Phys. **181**, 772–802 (2020) Zbl 1458.60050 MR 4160910

[3] Baur, E., Bertoin, J.: Cutting edges at random in large recursive trees. In: Stochastic analysis and applications 2014, Springer Proc. Math. Statist. 100, Springer, Cham, 51–76 (2014) Zbl 1390.60043 MR 3332709

[4] Baur, E., Bertoin, J.: Elephant random walks and their connection to Pólya-type urns. Phys. Rev. E **94**, art. 052134, 6 pp. (2016)

[5] Bercu, B.: A martingale approach for the elephant random walk. J. Phys. A **51**, art. 015201, 16 pp. (2018) Zbl 1392.60038 MR 3741953

[6] Bercu, B., Laulin, L.: On the center of mass of the elephant random walk. Stochastic Process. Appl. **133**, 111–128 (2021) Zbl 1469.60135 MR 4187306

[7] Bertenghi, M.: Functional limit theorems for the multi-dimensional elephant random walk. Stoch. Models **38**, 37–50 (2022) Zbl 1490.60105 MR 4359299

[8] Bertenghi, M.: Asymptotic normality of superdiffusive step-reinforced random walks. arXiv:2101.00906 (2021)

[9] Bertoin, J.: Noise reinforcement for Lévy processes. Ann. Inst. Henri Poincaré Probab. Statist. **56**, 2236–2252 (2020) Zbl 1477.60069 MR 4116724

[10] Bertoin, J.: Scaling exponents of step-reinforced random walks. Probab. Theory Related Fields **179**, 295–315 (2021) Zbl 1483.60065 MR 4221659

[11] Bertoin, J.: Universality of noise reinforced Brownian motions. In: In and out of equilibrium 3: Celebrating Vladas Sidoravicius, Progr. Probab. 77, Birkhäuser/Springer, Cham, 147–161 (2021) Zbl 1469.60253 MR 4237267

[12] Bingham, N. H., Goldie, C. M., Teugels, J. L.: Regular variation. Encyclopedia of Mathematics and its Applications 27, Cambridge University Press, Cambridge (1987) Zbl 0617.26001 MR 898871

[13] Bollobás, B., Riordan, O., Spencer, J., Tusnády, G.: The degree sequence of a scale-free random graph process. Random Structures Algorithms **18**, 279–290 (2001) Zbl 0985.05047 MR 1824277

[14] Businger, S.: The shark random swim (Lévy flight with memory). J. Statist. Phys. **172**, 701–717 (2018) Zbl 1400.82097 MR 3827299

[15] Coletti, C. F., Gava, R., Schütz, G. M.: Central limit theorem and related results for the elephant random walk. J. Math. Phys. **58**, art. 053303, 8 pp. (2017) Zbl 1375.60086 MR 3652225

[16] Coletti, C. F., Gava, R., Schütz, G. M.: A strong invariance principle for the elephant random walk. J. Statist. Mech. Theory Exp. **2017**, art. 123207, 8 pp. (2017) Zbl 1457.82148 MR 3748931

[17] Drmota, M.: Random trees. Springer, Wien (2009) Zbl 1170.05022 MR 2484382

[18] Durrett, R.: Random graph dynamics. Cambridge Series in Statistical and Probabilistic Mathematics 20, Cambridge University Press, Cambridge (2007) Zbl 1116.05001 MR 2271734

[19] Gnedenko, B. V., Kolmogorov, A. N.: Limit distributions for sums of independent random variables. Rev. ed., Addison-Wesley, Reading, MA (1968) Zbl 0056.36001 MR 233400

[20] Gut, A., Stadtmüller, U.: The number of zeros in elephant random walks with delays. Statist. Probab. Lett. **174**, art. 109112, 9 pp. (2021) Zbl 1478.60089 MR 4246204

[21] Ibragimov, I. A., Linnik, Y. V.: Independent and stationary sequences of random variables. Wolters-Noordhoff, Groningen (1971) Zbl 0219.60027 MR 322926

[22] Kubota, N., Takei, M.: Gaussian fluctuation for superdiffusive elephant random walks. J. Statist. Phys. **177**, 1157–1171 (2019) Zbl 1439.60043 MR 4034803

[23] Kürsten, R.: Random recursive trees and the elephant random walk. Phys. Rev. E **93**, art. 032111, 11 pp. (2016) MR 3652690

[24] Mahmoud, H. M.: Pólya urn models. Texts in Statistical Science Series, CRC Press, Boca Raton, FL (2009) Zbl 1149.60005 MR 2435823

[25] Meir, A., Moon, J. W.: Cutting down recursive trees. Math. Biosci. **21**, 173–181 (1974) Zbl 0288.05102

[26] Miyazaki, T., Takei, M.: Limit theorems for the 'laziest' minimal random walk model of elephant type. J. Statist. Phys. **181**, 587–602 (2020) Zbl 1460.60036 MR 4143637

[27] Najock, D., Heyde, C. C.: On the number of terminal vertices in certain random trees with an application to stemma construction in philology. J. Appl. Probab. **19**, 675–680 (1982) Zbl 0487.60012 MR 664852

[28] Pachon, A., Polito, F., Sacerdote, L.: Random graphs associated to some discrete and continuous time preferential attachment models. J. Statist. Phys. **162**, 1608–1638 (2016) Zbl 1336.05118 MR 3463790

[29] Pemantle, R.: A survey of random processes with reinforcement. Probab. Surv. **4**, 1–79 (2007) Zbl 1189.60138 MR 2282181

[30] Petersen, T. K.: Eulerian numbers. Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser/Springer, New York (2015) MR 3408615

[31] Schütz, G. M., Trimper, S.: Elephants can always remember: Exact long-range memory effects in a non-Markovian random walk. Phys. Rev. E **70**, art. 045101 (2004)

[32] Simon, H. A.: On a class of skew distribution functions. Biometrika **42**, 425–440 (1955) Zbl 0066.11201 MR 73085

[33] Stanley, R. P.: Enumerative combinatorics. Vol. 1. Cambridge Studies in Advanced Mathematics 49, Cambridge University Press, Cambridge (1997) Zbl 0889.05001 MR 1442260

[34] Tanny, S.: A probabilistic interpretation of Eulerian numbers. Duke Math. J. **40**, 717–722 (1973) Zbl 0284.05006 MR 340045