Electronic Journal of Probability

Electron. J. Probab. **21** (2016), no. 43, 1–23.
ISSN: 1083-6489 DOI: 10.1214/16-EJP4492

From infinite urn schemes to decompositions of self-similar Gaussian processes

Olivier Durieu[*] Yizao Wang[†]

Abstract

We investigate a special case of infinite urn schemes first considered by Karlin (1967), especially its occupancy and odd-occupancy processes. We first propose a natural randomization of these two processes and their decompositions. We then establish functional central limit theorems, showing that each randomized process and its components converge jointly to a decomposition of a certain self-similar Gaussian process. In particular, the randomized occupancy process and its components converge jointly to a decomposition of a time-changed Brownian motion $\mathbb{B}(t^\alpha), \alpha \in (0,1)$, and the randomized odd-occupancy process and its components converge jointly to a decomposition of a fractional Brownian motion with Hurst index $H \in (0,1/2)$. The decomposition in the latter case is a special case of the decomposition of bi-fractional Brownian motions recently investigated by Lei and Nualart (2009). The randomized odd-occupancy process can also be viewed as a correlated random walk, and in particular as a complement to the model recently introduced by Hammond and Sheffield (2013) as discrete analogues of fractional Brownian motions.

Keywords: infinite urn scheme; regular variation; functional central limit theorem; self-similar process; fractional Brownian motion; bi-fractional Brownian motion; decomposition; symmetrization.
AMS MSC 2010: Primary 60F17; 60G22, Secondary 60G15; 60G18.
Submitted to EJP on August 19, 2015, final version accepted on July 4, 2016.
Supersedes arXiv:1508.01506.
Supersedes HAL:hal-01184411.

1 Introduction

We consider the classical infinite urn scheme, sometimes referred to as the balls-in-boxes scheme. Namely, consider an infinite number of boxes labeled by $\mathbb{N} := \{1, 2, \ldots\}$, and suppose all boxes are empty at the beginning. Then, each round a ball is put into a

[*] Laboratoire de Mathématiques et Physique Théorique (UMR-CNRS 7350), Fédération Denis Poisson (FR-CNRS 2964), Université François–Rabelais de Tours, Parc de Grandmont, 37200 Tours, France.
E-mail: olivier.durieu@lmpt.univ-tours.fr

[†] Department of Mathematical Sciences, University of Cincinnati, 2815 Commons Way, Cincinnati, OH, 45221-0025, USA.
E-mail: yizao.wang@uc.edu

box with a random label sampled from a fixed distribution μ on $\mathbb{N}$, and the samplings of labels at different rounds are independent. This model has a very long history, dating back to at least Bahadur [1]. For a recent survey from the probabilistic point of view, see Gnedin et al. [11]. In particular, the sampling of the boxes forms naturally an exchangeable random partition of $\mathbb{N}$. Exchangeable random partitions have been extensively studied in the literature, and have connections to various areas in probability theory and related fields. See the nice monograph by Pitman [22] on random partitions and more general combinatorial stochastic processes. For various applications of the infinite urn schemes in biology, ecology, computational linguistics, among others, see for example Bunge and Fitzpatrick [6].

In this paper, we are interested in a specific infinite urn scheme. More precisely, we consider a probability measure μ on $\mathbb{N}$ satisfying a certain regular variation assumption with index $\alpha \in (0,1)$, to be defined in Section 2.1. This model was first considered by Karlin [14] and we will refer to it as the *Karlin model* in the rest of the paper.

We start by recalling the main results of Karlin [14]. Let $(Y_i)_{i\geq 1}$ represent the independent sampling from μ in each round $i \geq 1$, and

$$Y_{n,k} := \sum_{i=1}^{n} \mathbb{1}_{\{Y_i = k\}}, \quad n \geq 1,\, k \geq 1,$$

be the total counts of the label k sampled in the first n rounds, or equivalently the number of balls thrown into the box k in the first n rounds. In particular, Karlin investigated the asymptotics of two statistics: the total number of boxes that have been chosen in the first n rounds, denoted by

$$Z^*(n) := \sum_{k\geq 1} \mathbb{1}_{\{Y_{n,k} \neq 0\}},$$

and the total number of boxes that have been chosen by an odd number of times in the first n rounds, denoted by

$$U^*(n) := \sum_{k\geq 1} \mathbb{1}_{\{Y_{n,k} \text{ is odd}\}}.$$

The processes Z^* and U^* are referred to as the *occupancy process* and the *odd-occupancy process*, respectively. While Z^* is a natural statistics to consider in view of sampling different species, the investigation of U^* is motivated via the following light-bulb-switching point of view from Spitzer [26]. Each box k may represent the status (on/off) of a light bulb, and each time when k is sampled, the status of the corresponding light bulb is switched either from on to off or from off to on. In this way, assuming that all the light bulbs are off at the beginning, $U^*(n)$ represents the total number of light bulbs that are on at time n.

Central limit theorems have been established for both processes in [14], in the form of

$$\frac{Z^*(n) - \mathbb{E}Z^*(n)}{\sigma_n} \Rightarrow \mathcal{N}(0, \sigma_Z^2) \quad \text{and} \quad \frac{U^*(n) - \mathbb{E}U^*(n)}{\sigma_n} \Rightarrow \mathcal{N}(0, \sigma_U^2) \tag{1.1}$$

for some normalization σ_n, with σ_Z^2 and σ_U^2 explicitly given as the variances of the limiting normal distributions, and where $\Rightarrow$ denotes convergence in distribution. We remark that σ_n^2 is of the order n^α, up to a slowly varying function at infinity.

The next seemingly obvious task is to establish the functional central limit theorems for the two statistics. However, to the best of our knowledge, this has not been addressed in the literature. Here, by functional central limit theorems we are thinking of results in the form of (in terms of Z^*)

$$\left(\frac{Z^*(\lfloor nt \rfloor) - \mathbb{E}Z^*(\lfloor nt \rfloor)}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{Z}^*(t))_{t\in[0,1]}, \tag{1.2}$$

in the space $D([0,1])$ for some normalization sequence σ_n and a Gaussian process $\mathbb{Z}^*$. In view of (1.1) and the fact that σ_n^2 has the same order as n^α, the scaling limit $\mathbb{Z}^*$, if exists, is necessarily self-similar with index $\alpha/2$.

In this paper, instead of addressing only this question, we consider a more general framework by introducing a randomization of the Karlin model that consists in attaching independent Rademacher random variables to the boxes (see Section 2.1 for the exact definitions). The randomization of the Karlin model reveals certain rich structure of the model. In particular, it has a natural decomposition. Take the randomized occupancy process Z^ε for example. We will write

$$Z^\varepsilon(n) = Z_1^\varepsilon(n) + Z_2^\varepsilon(n)$$

and prove a joint weak convergence result in form of

$$\frac{1}{\sigma_n}(Z_1^\varepsilon(\lfloor nt\rfloor), Z_2^\varepsilon(\lfloor nt\rfloor), Z^\varepsilon(\lfloor nt\rfloor))_{t\in[0,1]} \Rightarrow (\mathbb{Z}_1(t), \mathbb{Z}_2(t), \mathbb{Z}(t))_{t\in[0,1]}\,,$$

in $D([0,1])^3$, such that

$$\mathbb{Z} = \mathbb{Z}_1 + \mathbb{Z}_2 \quad \text{with} \quad \mathbb{Z}_1 \text{ and } \mathbb{Z}_2 \text{ independent.}$$

In other words, the limit trivariate Gaussian process $(\mathbb{Z}_1(t), \mathbb{Z}_2(t), \mathbb{Z}(t))_{t\in[0,1]}$ can be constructed by first considering two independent Gaussian processes $\mathbb{Z}_1$ and $\mathbb{Z}_2$ with covariance to be specified, and then setting $\mathbb{Z}(t) := \mathbb{Z}_1(t) + \mathbb{Z}_2(t), t \in [0,1]$; in this way its finite-dimensional distributions are also determined. We refer to such results as *weak convergence to the decomposition of a Gaussian process*. Similar results for the randomized odd-occupancy process are also obtained. Here is a brief summary of the main results of the paper.

- As expected, various self-similar Gaussian processes appear in the limit. In this way, the randomized Karlin model and its components, including Z^* and U^* as special quenched cases, provide discrete counterparts of several self-similar Gaussian processes. These processes include notably the fractional Brownian motion with Hurst index $H = \alpha/2$, the bi-fractional Brownian motion with parameter $H = 1/2, K = \alpha$, and a new self-similar process $\mathbb{Z}_1$.
- Moreover, in view of the weak convergence to the decomposition, the randomized Karlin model are discrete counterparts of certain decompositions of self-similar Gaussian processes. The randomized occupancy process and its two components converge weakly to a new decomposition of the time-changed Brownian motion $(\mathbb{B}(t^\alpha))_{t\geq 0}, \alpha \in (0,1)$ (Theorem 2.1). The randomized odd-occupancy process and its two components converge weakly to a decomposition of the fractional Brownian motion with Hurst index $H = \alpha/2 \in (0,1/2)$ (Theorem 2.2). This decomposition is a particular case of the decompositions of bi-fractional Brownian motion recently discovered by Lei and Nualart [17].

Self-similar processes have been extensively studied in probability theory and related fields [9], often related to the notion of long-range dependence [21, 25]. Among the self-similar processes arising in the limit in this paper, the most widely studied one is the fractional Brownian motion. Fractional Brownian motions, as generalizations of Brownian motions, have been widely studied and used in various areas of probability theory and applications. These processes are the only centered Gaussian processes that are self-similar with stationary increments. The investigation of fractional Brownian motions dates back to Kolmogorov [16] and Mandelbrot and Van Ness [18]. As for limit theorems, there are already several models that converge to fractional Brownian motions in the literature. See [7, 10, 12, 15, 19, 20, 27] for a few representative examples. A more detailed and extensive survey of various models can be found in Pipiras and Taqqu [21]. Besides, we also obtain limit theorems for bi-fractional Brownian motions introduced by Houdré and Villa [13]. They often show up in decompositions of self-similar Gaussian processes; see for example [17, 24]. However, we do not find other discrete models for the bi-fractional Brownian motions in the literature. As for limit theorems illustrating decompositions of Gaussian processes as ours, we also find few examples in the literature; see Remark 2.6.

Our results connect the Karlin model, a discrete-time stochastic process, to several continuous-time self-similar Gaussian processes and their decompositions. By introducing new discrete counterparts, we hope to improve our understanding of these Gaussian processes. In particular, the proposed randomized Karlin model can also be viewed as correlated random walks, in a sense complementing the recent model introduced by Hammond and Sheffield [12] that scales to fractional Brownian motions with Hurst index $H \in (1/2, 1)$. Here, the randomized odd-occupancy process (U^ε below) is defined in a similar manner, and scales to fractional Brownian motions with $H \in (0, 1/2)$.

The paper is organized as follows. Section 2 introduces the model in details and present the main results as well as several comments. The proofs are based on a Poissonization technique. Section 3 introduces and investigates the Poissonized models. The de-Poissonization is established in Section 4.

2 Randomization of Karlin model and main results

2.1 Karlin model and its randomization

We have introduced the original Karlin model in Section 1. Here, we specify the regular variation assumption. Recall that μ is the common distribution of the $(Y_i)_{i\geq 1}$ and set $p_k := \mu(\{k\})$ for $k \in \mathbb{N}$. We assume that $(p_k)_{k\geq 1}$ is non-increasing, and define the infinite counting measure ν on $[0, \infty)$ by

$$\nu(A) := \sum_{j\geq 1} \delta_{\frac{1}{p_j}}(A)$$

for any Borel set A of $[0, \infty)$, where δ_x is the Dirac mass at x. For all $t > 0$, set

$$\nu(t) := \nu([0, t]) = \max\{j \geq 1 \mid p_j \geq 1/t\}, \tag{2.1}$$

where $\max \emptyset = 0$. Following Karlin [14], the main assumption is that $\nu(t)$ is a regularly varying function at ∞ with index α in $(0, 1)$, that is for all $x > 0$, $\lim_{t\to\infty} \nu(tx)/\nu(t) = x^\alpha$, or equivalently

$$\nu(t) = t^\alpha L(t),\ t > 0, \tag{2.2}$$

where L is a slowly varying function as $t \to \infty$, i.e. for all $x > 0$, $\lim_{t\to\infty} L(tx)/L(t) = 1$. For the sake of simplicity, one can think of

$$p_k \underset{k\to\infty}{\sim} Ck^{-\frac{1}{\alpha}} \text{ for some } \alpha \in (0, 1) \text{ and a normalizing constant } C > 0.$$

In this case, $\nu(t) \underset{t\to\infty}{\sim} C^\alpha t^\alpha$.

We have introduced two random processes considered in Karlin [14]: the occupancy process and the odd-occupancy process as

$$Z^*(n) := \sum_{k\geq 1} \mathbb{1}_{\{Y_{n,k}\neq 0\}} \quad \text{and} \quad U^*(n) := \sum_{k\geq 1} \mathbb{1}_{\{Y_{n,k} \text{ is odd}\}},$$

respectively. To introduce the randomization, let $\varepsilon := (\varepsilon_k)_{k\geq 1}$ be a sequence of i.i.d. Rademacher random variables (i.e. $\mathbb{P}(\varepsilon_k = 1) = \mathbb{P}(\varepsilon_k = -1) = 1/2$) defined on the same probability space as the $(Y_n)_{n\geq 1}$ and independent of them. In the sequel, we just say that ε is a Rademacher sequence in this situation, and implicitly ε is always assumed independent from $(Y_n)_{n\geq 1}$.

Now we introduce the *randomized occupancy process* and the *randomized odd-occupancy process* defined by

$$Z^\varepsilon(n) := \sum_{k\geq 1} \varepsilon_k \mathbb{1}_{\{Y_{n,k}\neq 0\}} \quad \text{and} \quad U^\varepsilon(n) := \sum_{k\geq 1} \varepsilon_k \mathbb{1}_{\{Y_{n,k} \text{ is odd}\}},$$

respectively. We actually will work with decompositions of these two processes given by

$$Z^\varepsilon(n) = Z_1^\varepsilon(n) + Z_2^\varepsilon(n) \quad \text{and} \quad U^\varepsilon(n) = U_1^\varepsilon(n) + U_2^\varepsilon(n),$$

where

$$Z_1^\varepsilon(n) := \sum_{k\geq 1} \varepsilon_k \left(\mathbb{1}_{\{Y_{n,k}\neq 0\}} - p_k(n)\right) \quad \text{and} \quad Z_2^\varepsilon(n) := \sum_{k\geq 1} \varepsilon_k p_k(n), \quad n \geq 1, \tag{2.3}$$

$$U_1^\varepsilon(n) := \sum_{k\geq 1} \varepsilon_k \left(\mathbb{1}_{\{Y_{n,k} \text{ is odd}\}} - q_k(n)\right) \quad \text{and} \quad U_2^\varepsilon(n) := \sum_{k\geq 1} \varepsilon_k q_k(n), \quad n \geq 1, \tag{2.4}$$

with for all $k \geq 1$ and $n \geq 1$,

$$p_k(n) := \mathbb{P}\left(Y_{n,k} \neq 0\right) = 1 - (1-p_k)^n,$$
$$q_k(n) := \mathbb{P}\left(Y_{n,k} \text{ is odd}\right) = \frac{1}{2}(1 - (1-2p_k)^n).$$

In the preceding definitions, the exponent ε refers to the randomness given by the Rademacher sequence $(\varepsilon_k)_{k\geq 1}$. Nevertheless, in some of the following statements, the sequence of $(\varepsilon_k)_{k\geq 1}$ can be chosen fixed (deterministic) in $\{-1,1\}^{\mathbb{N}}$. Then the corresponding processes can be considered as "quenched" versions of the randomized processes. For this purpose, it is natural to introduce the centering with $p_k(n)$ and $q_k(n)$ respectively above. Actually, we will establish quenched weak convergence for Z_1^ε and U_1^ε (see Theorem 2.3 and Remark 2.4). With a little abuse of language, for both cases we keep ε in the notation and add an explanation like '*for a Rademacher sequence* ε' or '*for all fixed* $\varepsilon \in \{-1,1\}^{\mathbb{N}}$', respectively.

2.2 Main results

As mentioned in the introduction, we are interested in the scaling limits of the previously defined processes. We denote by $D([0,1])$ the Skorohod space of cadlag functions on $[0,1]$ with the Skorohod topology (see [2]). Throughout, we write

$$\sigma_n := n^{\alpha/2} L(n)^{1/2},$$

where α and L are the same as in the regular variation assumption (2.2). Observe that $\nu(n) = L(n) = \sigma_n = 0$ for $n < 1/p_1$. Therefore, when writing $1/\sigma_n$ we always assume implicitly $n \geq 1/p_1$. Below are the main results of this paper.

Theorem 2.1. *For a Rademacher sequence* ε,

$$\frac{1}{\sigma_n}\left(Z_1^\varepsilon(\lfloor nt\rfloor), Z_2^\varepsilon(\lfloor nt\rfloor), Z^\varepsilon(\lfloor nt\rfloor)\right)_{t\in[0,1]} \Rightarrow \left(\mathbb{Z}_1(t), \mathbb{Z}_2(t), \mathbb{Z}(t)\right)_{t\in[0,1]},$$

in $(D([0,1]))^3$, *where* $\mathbb{Z}_1, \mathbb{Z}_2, \mathbb{Z}$ *are centered Gaussian processes, such that*

$$\mathbb{Z} = \mathbb{Z}_1 + \mathbb{Z}_2,$$

$\mathbb{Z}_1$ and $\mathbb{Z}_2$ are independent, and they have covariances

$$\begin{aligned}\operatorname{Cov}(\mathbb{Z}_1(s),\mathbb{Z}_1(t)) &= \Gamma(1-\alpha)\left((s+t)^\alpha - \max(s,t)^\alpha\right),\\ \operatorname{Cov}(\mathbb{Z}_2(s),\mathbb{Z}_2(t)) &= \Gamma(1-\alpha)\left(s^\alpha + t^\alpha - (s+t)^\alpha\right),\\ \operatorname{Cov}(\mathbb{Z}(s),\mathbb{Z}(t)) &= \Gamma(1-\alpha)\min(s,t)^\alpha, \quad s,t\ge 0.\end{aligned}$$

Theorem 2.2. *For a Rademacher sequence ε,*

$$\frac{1}{\sigma_n}\left(U_1^\varepsilon(\lfloor nt\rfloor), U_2^\varepsilon(\lfloor nt\rfloor), U^\varepsilon(\lfloor nt\rfloor)\right)_{t\in[0,1]} \Rightarrow \left(\mathbb{U}_1(t),\mathbb{U}_2(t),\mathbb{U}(t)\right)_{t\in[0,1]},$$

in $(D([0,1]))^3$, where $\mathbb{U}_1,\mathbb{U}_2,\mathbb{U}$ are centered Gaussian processes such that

$$\mathbb{U} = \mathbb{U}_1 + \mathbb{U}_2,$$

$\mathbb{U}_1$ and $\mathbb{U}_2$ are independent, and they have covariances

$$\begin{aligned}\operatorname{Cov}(\mathbb{U}_1(s),\mathbb{U}_1(t)) &= \Gamma(1-\alpha)2^{\alpha-2}\left((s+t)^\alpha - |t-s|^\alpha\right),\\ \operatorname{Cov}(\mathbb{U}_2(s),\mathbb{U}_2(t)) &= \Gamma(1-\alpha)2^{\alpha-2}\left(s^\alpha + t^\alpha - (s+t)^\alpha\right),\\ \operatorname{Cov}(\mathbb{U}(s),\mathbb{U}(t)) &= \Gamma(1-\alpha)2^{\alpha-2}\left(s^\alpha + t^\alpha - |t-s|^\alpha\right), \quad s,t\ge 0.\end{aligned}$$

To achieve these results, we will first prove the convergence of the first (Z_1^ε and U_1^ε) and the second (Z_2^ε and U_2^ε) components, respectively. For the first components we have the following stronger result.

Theorem 2.3. *For all fixed $\varepsilon\in\{-1,1\}^{\mathbb{N}}$,*

$$\left(\frac{Z_1^\varepsilon(\lfloor nt\rfloor)}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{Z}_1(t))_{t\in[0,1]} \quad \text{and} \quad \left(\frac{U_1^\varepsilon(\lfloor nt\rfloor)}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{U}_1(t))_{t\in[0,1]},$$

in $D([0,1])$, where $\mathbb{Z}_1$ and $\mathbb{U}_1$ are as in Theorems 2.1 and 2.2.

Remark 2.4. Theorem 2.3 is a quenched functional central limit theorem. In particular, when taking $\varepsilon = \vec{1} = (1,1,\dots)$, Theorem 2.3 gives functional versions of the central limit theorems for $Z^*(n)$ and $U^*(n)$ established in Karlin [14] (formally stated in (1.1)): the (non-randomized) occupancy and odd-occupancy processes of the Karlin model scale to the continuous-time processes $\mathbb{Z}_1$ and $\mathbb{U}_1$, respectively. Moreover, as the limits in Theorem 2.3 do not depend on the value of ε, this implies the annealed functional central limit theorems (the same statement of Theorem 2.3 remains true for a Rademacher sequence ε).

Now we take a closer look at the processes appearing in Theorem 2.1 and Theorem 2.2 and the corresponding decompositions. The decomposition of $\mathbb{U}$ is a special case of the general decompositions established in Lei and Nualart [17] for bi-fractional Brownian motions. Recall that a bi-fractional Brownian motion with parameter $H\in(0,1), K\in(0,1]$ is a centered Gaussian process with covariance function

$$R^{H,K}(s,t) = \frac{1}{2^K}\left(\left(t^{2H}+s^{2H}\right)^K - |t-s|^{2HK}\right). \tag{2.5}$$

The case $K=1$ corresponds to the fractional Brownian motion with Hurst index H. It is noticed in [17] that one can write

$$\frac{1}{2^K}\left(t^{2HK}+s^{2HK}-|t-s|^{2HK}\right) = R^{H,K}(s,t) + \frac{1}{2^K}\left(t^{2HK}+s^{2HK}-(t^{2H}+s^{2H})^K\right), \tag{2.6}$$

where the left-hand side above is a multiple of the covariance function of a fractional Brownian motion with Hurst index HK, and the second term in the right-hand side above is positive-definite and hence a covariance function. Therefore, (2.6) induces a decomposition of a fractional Brownian motion with Hurst index HK into a bi-fractional Brownian motion and another self-similar Gaussian process.

Comparing this to Theorem 2.2, we notice that our decomposition of $\mathbb{U}$ corresponds to the special case of (2.6) with $H = 1/2, K = \alpha$. Up to a multiplicative constant, $\mathbb{U}$ is a fractional Brownian motion with Hurst index $H = \alpha/2$. The process $\mathbb{U}_1$ is the bi-fractional Brownian motion with $H = 1/2, K = \alpha$, and it is also known as the odd-part of the two-sided fractional Brownian motion; see Dzhaparidze and van Zanten [8]. That is

$$(\mathbb{U}_1(t))_{t\geq 0} \stackrel{fdd}{=} \sqrt{2^\alpha\Gamma(1-\alpha)}\left(\frac{1}{2}(\mathbb{B}^{\alpha/2}(t) - \mathbb{B}^{\alpha/2}(-t))\right)_{t\geq 0},$$

where $\mathbb{B}^{\alpha/2}$ is a two-sided fractional Brownian motion on $\mathbb{R}$ with Hurst index $\alpha/2 \in (0,1)$. The process $\mathbb{U}_2$ admits a representation

$$\mathbb{U}_2(t) = 2^{\alpha/2-1}\sqrt{\alpha}\int_0^\infty (1-e^{st})s^{-\frac{\alpha+1}{2}}d\mathbb{B}(s),\ t>0,$$

where $(\mathbb{B}(t))_{t\in[0,1]}$ is the standard Brownian motion. It is shown that $\mathbb{U}_2(t)$ has a version with infinitely differentiable path for $t \in (0,\infty)$ and absolutely continuous path for $t \in [0,\infty)$. At the same time, $\mathbb{U}_2$ also appears in the decomposition of sub-fractional Brownian motions [4, 24].

For the decomposition of $\mathbb{Z}$ in Theorem 2.1, to the best of our knowledge it is new in the literature. Remark that $\mathbb{Z}$ is simply a time-changed Brownian motion $(\mathbb{Z}(t))_{t\geq 0} \stackrel{fdd}{=} \Gamma(1-\alpha)(\mathbb{B}(t^\alpha))_{t\geq 0}$, and that $\mathbb{Z}_2 \stackrel{fdd}{=} 2^{-\alpha/2+1}\mathbb{U}_2$. The latter is not surprising as the coefficients $q_k(n)$ and $p_k(n)$ have the same asymptotic behavior. However, we cannot find related reference for $\mathbb{Z}_1$ in the literature. The following remark on $\mathbb{Z}_1$ has its own interest.

Remark 2.5. The process $\mathbb{Z}_1$ may be related to bi-fractional Brownian motions as follows. One can write

$$(s^{1/\alpha}+t^{1/\alpha})^\alpha - |s-t| = 2\left[\left(s^{1/\alpha}+t^{1/\alpha}\right)^\alpha - \max(s,t)\right] + \left[s+t-\left(s^{1/\alpha}+t^{1/\alpha}\right)^\alpha\right], s,t\geq 0.$$

That is,

$$(\mathbb{V}(t))_{t\geq 0} \stackrel{fdd}{=} \left(2\mathbb{Z}_1(t^{1/\alpha}) + \mathbb{Z}_2(t^{1/\alpha})\right)_{t\geq 0},$$

where $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are as before and independent, and $\mathbb{V}$ is a centered Gaussian process with covariance

$$\mathrm{Cov}(\mathbb{V}(s),\mathbb{V}(t)) = \Gamma(1-\alpha)2^\alpha R^{1/(2\alpha),\alpha}(s,t).$$

Therefore, as another consequence of our results, we have shown that for the bi-fractional Brownian motions, the covariance function $R^{H,K}$ in (2.5) is well defined for $H = 1/(2\alpha), K = \alpha$ for all $\alpha \in (0,1)$. The range $\alpha \in (0,1/2]$ is new.

Remark 2.6. We are not aware of other limit theorems for the decomposition of processes in a similar manner as ours, but with two exceptions. One is the symmetrization well investigated in the literature of empirical processes [28]. Take for a simple example the empirical distribution function

$$\mathbb{F}_n(t) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i\leq t\}}$$

where $X_1, X_2, \ldots$ are i.i.d. with uniform $(0,1)$ distribution. By symmetrization one considers an independent Rademacher sequence ε and

$$\mathbb{F}_n^\varepsilon(t) := \frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{1}_{\{X_i\leq t\}},\quad \mathbb{F}_n^{\varepsilon,1}(t) := \frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(\mathbb{1}_{\{X_i\leq t\}} - t\right)\quad \text{and}\quad \mathbb{F}_n^{\varepsilon,2}(t) := \frac{1}{n}\sum_{i=1}^n \varepsilon_i t.$$

It is straight-forward to establish

$$\sqrt{n}\,(\mathbb{F}_n^{\varepsilon}(t), \mathbb{F}_n^{\varepsilon,1}(t), \mathbb{F}_n^{\varepsilon,2}(t))_{t\in[0,1]} \Rightarrow (\mathbb{B}(t), \mathbb{B}(t) - t\mathbb{B}(1), t\mathbb{B}(1))_{t\in[0,1]}.$$

This provides an interpretation of the definition of Brownian bridge via $\mathbb{B}^{bridge}(t) := \mathbb{B}(t) - t\mathbb{B}(1), t \in [0,1]$. The other example of limit theorems for decompositions is the recent paper by Bojdecki and Talarczyk [5] who provided a particle-system point of view for the decomposition of fractional Brownian motions. The model considered there is very different from ours, and so is the decomposition in the limit.

To prove the convergence of each individual process, we apply a Poissonization technique which was already used by Karlin [14]. Each of the Poissonized processes $\tilde{\mathbb{Z}}_1^{\varepsilon}, \tilde{\mathbb{Z}}_2^{\varepsilon}, \tilde{\mathbb{U}}_1^{\varepsilon}, \tilde{\mathbb{U}}_2^{\varepsilon}$ is an infinite sum of independent random variables of which the covariances are easy to calculate, and thus the finite-dimensional convergence follows immediately. This finite-dimensional convergence is already new comparing to [14] but it does not involve any new technique. A first challenging question for us is to establish the tightness for $\tilde{Z}_1^{\varepsilon}$ and $\tilde{U}_1^{\varepsilon}$. Karlin [14] did not consider the functional central limit theorems, and in particular to obtain the tightness one needs to work harder. For this purpose we apply a chaining argument. Once the weak convergence for the Poissonized process is established, we couple the Poissonized process with the original one and bound the difference. The second technical challenge lies in this de-Poissonization step. Again, our de-Poissonization lemmas are more involved than in [14] since we work with $D([0,1])$-valued random variables.

Remark 2.7. One can prove the weak convergences $(Z^{\varepsilon}(\lfloor nt \rfloor)/\sigma_n)_{t\in[0,1]} \Rightarrow (\mathbb{Z}(t))_{t\in[0,1]}$ and $(U^{\varepsilon}(\lfloor nt \rfloor)/\sigma_n)_{t\in[0,1]} \Rightarrow (\mathbb{U}(t))_{t\in[0,1]}$ directly, without using the decomposition. We do not present the proofs here as they do not provide insights on the decompositions of the limiting processes. Nevertheless, the later convergence has its own interest as explained in the next section.

2.3 Correlated random walks

Another motivation for this paper is to give a model of correlated random walks complementing a model proposed by Hammond and Sheffield [12] as a discrete counterpart to the fractional Brownian motion. Here we focus our discussion on the process U^{ε} and first explain that it can be interpreted as a correlated random walk by writing

$$U^{\varepsilon}(n) = X_1 + \cdots + X_n, \tag{2.7}$$

where the steps $(X_i)_{i\geq 1}$ are random variables taking values in $\{-1, 1\}$ with uniform probability. Here, unlike the usual random walks, the steps are dependent and the dependence is determined by the random partition of $\mathbb{N}$ generated by our balls-in-boxes scheme.

To obtain the representation of U^{ε} in (2.7), consider the sequence $(Y_n)_{n\geq 1}$ of independent copies with law μ and the random partition of $\mathbb{N}$ induced by the equivalence relation $i \sim j$ if and only if $Y_i = Y_j$. That is, the integers i and j are in the same component of the partition if and only if the i-th and j-th balls fall in the same box. Once the sequence $(Y_n)_{n\geq 1}$ is given and thus all components are determined, one can define the steps $(X_i)_{i\geq 1}$ as follows: Consider the sequence of independent Rademacher random variables $(\varepsilon_k)_{k\geq 1}$ (also independent of $(Y_n)_{n\geq 1}$). For each $k \geq 1$, list all the elements in the component k (defined as $\{i \in \mathbb{N} : Y_i = k\}$) in increasing order $i_1 < i_2 < \cdots$, and set $X_{i_1} := \varepsilon_k$ and iteratively $X_{i_{\ell+1}} := -X_{i_\ell}$, $\ell \geq 1$. In this way, it is easy to see that each X_i is taking values -1 or 1 with equal probabilities and that, conditioning on $(Y_n)_{n\geq 1}$, X_i and X_j are completely dependent if $i \sim j$ whereas they are independent if $i \not\sim j$. Further, for the m first integers in the component k, the corresponding sum $X_{i_1} + \cdots + X_{i_m}$ equals to ε_k if m is odd and vanishes if m is even. The verification of (2.7) is now straight-forward.

The above discussion describes how to construct correlated random walks from random partitions in two steps. The first is to sample the random partition. The second is to assign ± 1 values to $(X_i)_{i \geq 1}$ conditioned on the sampled random partition. The motivation for this discussion comes from a similar model of correlated random walk introduced by Hammond and Sheffield [12]. Hammond and Sheffield also constructed a collection of random variables taking values in $\{-1, 1\}$ for which the dependence among them is determined by a random partition of $\mathbb{Z}$. In their model, the random partition is given by the (infinitely many) connected components of a random graph on $\mathbb{Z}$ which is constructed by linking each integer i to the integer $i - Z_i$, where the $(Z_i)_{i \in \mathbb{Z}}$ are positive i.i.d. random variables with distribution in the domain of attraction of an α_0-stable law for some $\alpha_0 \in (0, 1/2)$. The ± 1 values are assigned such that $X_i = X_j$ if i and j belong to the same component and they are independent otherwise. The main result of [12] was to prove that the scaling limit of this correlated random walk (under appropriate normalization) is a fractional Brownian motion with Hurst index $\alpha_0 + 1/2 \in (1/2, 1)$. This gives discrete counterparts to fractional Brownian motions with Hurst index greater than $1/2$.

Our correlated random walk $(U^\varepsilon(n))_{n \geq 1}$ (as described by (2.7)) thus gives a complementary model for fractional Brownian motions with Hurst index smaller than $1/2$ since, focusing on U^ε in Theorem 2.2, we have the following result.

Corollary 2.8. *Set* $\eta_n^2 := \Gamma(1-\alpha) 2^{\alpha-1} n^\alpha L(n)$*. The process* $(U^\varepsilon(\lfloor nt \rfloor)/\eta_n)_{t \in [0,1]}$ *converges in distribution, in* $D([0,1])$*, to a fractional Brownian motion with Hurst index* $\alpha/2 \in (0, 1/2)$*.*

There are two differences between the Hammond–Sheffield model and the randomized odd-occupancy process U^ε: first, the underlying random partition is different: notably, the random partition in the infinite urn scheme is exchangeable, while this is not the case for the random partition of $\mathbb{Z}$ introduced in [12]; rather, the random partition there inherits certain long-range dependence which essentially determines that the Hurst index in the limit must be in $(1/2, 1)$. Second, the ± 1 assigning rule is different since for the Hammond–Sheffield model all the random variables indexed in the same component take the *same value*. The alternative way of assigning the ± 1 by *alternating the values* along each component is the key idea in our framework. Actually, Hammond and Sheffield [12] suggested, as an open problem, to apply this alternative assigning rule to their model and asked whether the modified model scales to a fractional Brownian motion with Hurst index in $(0, 1/2)$. In our point of view, in order to obtain a discrete model in the similar flavor of the Hammond–Sheffield model that scales to a fractional Brownian motion with Hurst index $H \in (0, 1/2)$, the alternative assigning rule is crucial, while the underlying random graph with long memory is not that essential. Our results support this point of view. At the same time, the aforementioned suggestion in [12] remains a challenging model to analyze.

3 Poissonization

Recall that we are interested in the processes Z^ε and U^ε and in the decompositions $Z^\varepsilon = Z_1^\varepsilon + Z_2^\varepsilon$ and $U^\varepsilon = U_1^\varepsilon + U_2^\varepsilon$ as defined in (2.3) and (2.4).

3.1 Definitions and preliminary results

The first step in the proofs is to consider the Poissonized versions of all the preceding processes in order to deal with sums of independent variables. Let N be a Poisson process with intensity 1, independent of the sequence $(Y_n)_{n \geq 1}$ and of the Rademacher sequence ε considered before. We set

$$N_k(t) := \sum_{\ell=1}^{N(t)} \mathbb{1}_{\{Y_\ell = k\}}, \quad t \geq 0, k \geq 1.$$

Then the processes N_k, $k \geq 1$, are independent Poisson processes with respective intensity p_k. Now we consider the Poissonized processes, for all $t \geq 0$,

$$\tilde{Z}^\varepsilon(t) := \sum_{k\geq 1} \varepsilon_k \mathbb{1}_{\{N_k(t)\neq 0\}} \quad \text{and} \quad \tilde{U}^\varepsilon(t) := \sum_{k\geq 1} \varepsilon_k \mathbb{1}_{\{N_k(t) \text{ is odd}\}}.$$

These Poissonized randomized occupancy and odd-occupancy processes have similar decompositions as the original processes

$$\tilde{Z}^\varepsilon = \tilde{Z}_1^\varepsilon + \tilde{Z}_2^\varepsilon \quad \text{and} \quad \tilde{U}^\varepsilon = \tilde{U}_1^\varepsilon + \tilde{U}_2^\varepsilon$$

with

$$\begin{aligned}\tilde{Z}_1^\varepsilon(t) &:= \sum_{k\geq 1} \varepsilon_k \left(\mathbb{1}_{\{N_k(t)\neq 0\}} - \tilde{p}_k(t)\right), \quad \tilde{Z}_2^\varepsilon(t) := \sum_{k\geq 1} \varepsilon_k \tilde{p}_k(t),\\ \tilde{U}_1^\varepsilon(t) &:= \sum_{k\geq 1} \varepsilon_k \left(\mathbb{1}_{\{N_k(t) \text{ is odd}\}} - \tilde{q}_k(t)\right), \quad \tilde{U}_2^\varepsilon(t) := \sum_{k\geq 1} \varepsilon_k \tilde{q}_k(t),\end{aligned}$$

and

$$\begin{aligned}\tilde{p}_k(t) &:= \mathbb{P}(N_k(t) \neq 0) = 1 - e^{-p_k t},\\ \tilde{q}_k(t) &:= \mathbb{P}(N_k(t) \text{ is odd}) = \frac{1}{2}(1 - e^{-2p_k t}).\end{aligned}$$

Using the independence and the stationarity of the increments of Poisson processes, we derive the following useful identities. For all $0 \leq s \leq t$ and all $k \geq 1$,

$$0 \leq \tilde{p}_k(t) - \tilde{p}_k(s) = (1 - \tilde{p}_k(s))\tilde{p}_k(t-s) \leq \tilde{p}_k(t-s), \tag{3.1}$$

$$0 \leq \tilde{q}_k(t) - \tilde{q}_k(s) = (1 - 2\tilde{q}_k(s))\tilde{q}_k(t-s) \leq \tilde{q}_k(t-s). \tag{3.2}$$

Note that, in particular, the functions $\tilde{p}_k$ and $\tilde{q}_k$ are sub-additive. Further, we will have to deal with the asymptotics of the sums over k of the $\tilde{p}_k$ or $\tilde{q}_k$. For this purpose, recall that (see [14, Theorem 1]) the assumption (2.2) implies

$$V(t) := \sum_{k\geq 1} (1 - e^{-p_k t}) \sim \Gamma(1-\alpha) t^\alpha L(t), \ \text{ as } t \to \infty. \tag{3.3}$$

We will need a further estimate on the asymptotic of $V(t)$ that is stated in the following lemma.

Lemma 3.1. *For all $\gamma \in (0, \alpha)$, there exists a constant $C_\gamma > 0$ such that*

$$V(nt) \leq C_\gamma t^\gamma \sigma_n^2, \ \textit{ uniformly in } t \in [0,1], n \geq 1.$$

Proof. Recall the definition of the integer-valued function ν in (2.1). By integration by parts, we have for all $t > 0$,

$$V(t) = \int_0^\infty (1 - e^{-t/x}) d\nu(x) = \int_0^\infty x^{-2} e^{-1/x} \nu(tx) dx.$$

Observe that $\nu(t) = 0$ if and only if $t \in [0, 1/p_1)$ by definition, and in particular $L(t) = 0$ if and only if $t \in [0, 1/p_1)$. Thus,

$$\frac{V(nt)}{\sigma_n^2} = \int_{1/(ntp_1)}^\infty x^{-2} e^{-1/x} \nu(ntx) dx = t^\alpha \int_{1/(ntp_1)}^\infty x^{\alpha-2} e^{-1/x} \frac{L(ntx)}{L(n)} dx.$$

Now we introduce

$$L^*(t) = \begin{cases} L(1/p_1) & \text{if } t \in [0, 1/p_1) \\ L(t) & \text{if } t \in [1/p_1, \infty) \end{cases},$$

and obtain

$$\frac{V(nt)}{\sigma_n^2} \le t^\alpha \int_0^\infty x^{\alpha-2} e^{-1/x} \frac{L^*(ntx)}{L^*(n)} dx.$$

Let $\delta > 0$ be such that $\alpha + \delta < 1$ and $\alpha - \delta > \gamma$. Observe that L^* has the same asymptotic behavior as L by definition. In addition, L^* is bounded away from 0 and ∞ on any compact set of $[0, \infty)$. Thus, by Potter's theorem (see [3, Theorem 1.5.6]) there exists a constant $C_\delta > 0$ such that for all $x, y > 0$

$$\frac{L^*(x)}{L^*(y)} \le C_\delta \max\left(\left(\frac{x}{y}\right)^\delta, \left(\frac{x}{y}\right)^{-\delta}\right).$$

We infer, uniformly in $t \in [0, 1]$,

$$\begin{aligned}\frac{V(nt)}{\sigma_n^2} &\le C_\delta t^\alpha \int_0^\infty x^{\alpha-2} e^{-1/x} \max\left((tx)^\delta, (tx)^{-\delta}\right) dx \\ &\le C_\delta t^{\alpha-\delta} \left(\int_0^1 x^{\alpha-\delta-2} e^{-1/x} dx + \int_1^\infty x^{\alpha+\delta-2} e^{-1/x} dx\right),\end{aligned}$$

and both integrals are finite (the second one because we have taken δ such that $\alpha+\delta < 1$). Further, $t^{\alpha-\delta} \le t^\gamma$ for all $t \in [0, 1]$ and thus the lemma is proved. □

3.2 Functional central limit theorems

We now establish the invariance principles for the Poissonized processes.

Proposition 3.2. *For all fixed* $\varepsilon \in \{-1, 1\}^{\mathbb{N}}$,

$$\left(\frac{\tilde{Z}_1^\varepsilon(nt)}{\sigma_n}\right)_{t \in [0,1]} \Rightarrow (\mathbb{Z}_1(t))_{t \in [0,1]} \quad \text{and} \quad \left(\frac{\tilde{U}_1^\varepsilon(nt)}{\sigma_n}\right)_{t \in [0,1]} \Rightarrow (\mathbb{U}_1(t))_{t \in [0,1]},$$

in $D([0, 1])$*, where* $\mathbb{Z}_1$ *is as in Theorem 2.1 and* $\mathbb{U}_1$ *is as in Theorem 2.2.*

Proof. In the sequel $\varepsilon \in \{-1, 1\}^{\mathbb{N}}$ is fixed. The proof is divided into three steps.

(i) The covariances. Using the independence of the N_k, and that $\varepsilon_k^2 = 1$ for all $k \ge 1$, we infer that for all $0 \le s \le t$,

$$\begin{aligned}\operatorname{Cov}\left(\tilde{Z}_1^\varepsilon(ns), \tilde{Z}_1^\varepsilon(nt)\right) &= \sum_{k \ge 1} \left(\mathbb{P}(N_k(ns) \ne 0,\, N_k(nt) \ne 0) - \tilde{p}_k(ns)\tilde{p}_k(nt)\right) \\ &= \sum_{k \ge 1} \left((1 - e^{-p_k ns}) - (1 - e^{-p_k ns})(1 - e^{-p_k nt})\right) \\ &= V(n(s+t)) - V(nt),\end{aligned}$$

whence by (3.3),

$$\lim_{n \to \infty} \frac{1}{\sigma_n^2} \operatorname{Cov}\left(\tilde{Z}_1^\varepsilon(ns), \tilde{Z}_1^\varepsilon(nt)\right) = \Gamma(1 - \alpha)\left((s+t)^\alpha - t^\alpha\right).$$

For the odd-occupancy process, using the independence and the stationarity of the increments of the Poisson processes, for $0 \leq s \leq t$,

$$\begin{aligned}\operatorname{Cov}\left(\tilde{U}_1^\varepsilon(ns), \tilde{U}_1^\varepsilon(nt)\right) &= \sum_{k\geq 1} \left(\mathbb{P}(N_k(ns) \text{ is odd, } N_k(nt) \text{ is odd}) - \tilde{q}_k(ns)\tilde{q}_k(nt)\right) \\ &= \sum_{k\geq 1} \left(\tilde{q}_k(ns)(1 - \tilde{q}_k(n(t-s))) - \tilde{q}_k(ns)\tilde{q}_k(nt)\right) \\ &= \frac{1}{4}\sum_{k\geq 1}(1 - e^{-2p_k ns})(e^{-2p_k n(t-s)} + e^{-2p_k nt}) \\ &= \frac{1}{4}\left(V(2n(t+s)) - V(2n(t-s))\right).\end{aligned}$$

Thus, again by (3.3),

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \operatorname{Cov}\left(\tilde{U}_1^\varepsilon(ns), \tilde{U}_1^\varepsilon(nt)\right) = \Gamma(1-\alpha)2^{\alpha-2}\left((t+s)^\alpha - (t-s)^\alpha\right).$$

(ii) Finite-dimensional convergence. The finite-dimensional convergence for both processes is a consequence of the Lindeberg central limit theorem, using the Cramér–Wold device. Indeed, for any choice of constants $a_1, \ldots, a_d \in \mathbb{R}$, $d \geq 1$, and any reals $t_1, \ldots, t_d \in [0,1]$, the random variables

$$\varepsilon_k \sum_{i=1}^d a_i(\mathbb{1}_{\{N_k(nt_i)\neq 0\}} - \tilde{p}_k(nt_i)), \quad k \geq 1, n \geq 1$$

are independent and uniformly bounded. This entails the finite-dimensional convergence for $(\tilde{Z}_1^\varepsilon(nt)/\sigma_n)_{t\in[0,1]}$. The proof for $(\tilde{U}_1^\varepsilon(nt)/\sigma_n)_{t\in[0,1]}$ is similar.

(iii) Tightness. The proof of the tightness is technical and delayed to Section 3.3. □

Proposition 3.3. *For any Rademacher sequence $\varepsilon = (\varepsilon_k)_{k\geq 1}$,*

$$\left(\frac{\tilde{Z}_2^\varepsilon(nt)}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{Z}_2(t))_{t\in[0,1]} \quad \text{and} \quad \left(\frac{\tilde{U}_2^\varepsilon(nt)}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{U}_2(t))_{t\in[0,1]},$$

in $D([0,1])$, where $\mathbb{Z}_2$ is as in Theorem 2.1 and $\mathbb{U}_2$ is as in Theorem 2.2.

Proof. First remark that, since for all $t \geq 0$, $\tilde{q}_k(t) = \frac{1}{2}\tilde{p}_k(2t)$, we have $\tilde{U}_2^\varepsilon(t) = \frac{1}{2}\tilde{Z}_2^\varepsilon(2t)$. Thus the second convergence follows from the first one.

(i) The covariances. Since the ε_k are independent, using (3.3), we have for all $t, s \geq 0$,

$$\begin{aligned}\frac{1}{\sigma_n^2}\operatorname{Cov}(\tilde{Z}_2^\varepsilon(nt), \tilde{Z}_2^\varepsilon(ns)) &= \frac{1}{\sigma_n^2}\sum_{k\geq 1}\mathbb{E}(\varepsilon_k^2)\tilde{p}_k(nt)\tilde{p}_k(ns) \\ &= \frac{1}{\sigma_n^2}\sum_{k\geq 1}(1 - e^{-p_k nt})(1 - e^{-p_k ns}) \\ &= \frac{1}{\sigma_n^2}\left(V(nt) + V(ns) - V(n(t+s))\right) \\ &\longrightarrow \Gamma(1-\alpha)\left(t^\alpha + s^\alpha - (t+s)^\alpha\right) \text{ as } n \to \infty.\end{aligned}$$

(ii) Finite-dimensional convergence. Since $\tilde{Z}_2^\varepsilon$ is a sum of independent bounded random variables, the finite-dimensional convergence follows from the Cramér–Wold device and the Lindeberg central limit theorem.

(iii) Tightness. Let p be a positive integer. By Burkholder inequality, there exists a constant $C_p > 0$ such that for all $0 \le s \le t \le 1$,

$$\mathbb{E}\left|\frac{1}{\sigma_n}\left(\tilde{Z}_2^\varepsilon(nt) - \tilde{Z}_2^\varepsilon(ns)\right)\right|^{2p} \le C_p \frac{1}{\sigma_n^{2p}}\left(\sum_{k\ge1}(\tilde{p}_k(nt) - \tilde{p}_k(ns))^2\right)^p$$
$$\le C_p \frac{1}{\sigma_n^{2p}}\left(\sum_{k\ge1}\tilde{p}_k(n(t-s))^2\right)^p = C_p\left(\frac{V(n(t-s))}{\sigma_n^2}\right)^p.$$

Now we use Lemma 3.1. Let $\gamma \in (0,\alpha)$. There exists $C_\gamma > 0$ such that

$$\mathbb{E}\left|\frac{1}{\sigma_n}\left(\tilde{Z}_2^\varepsilon(nt) - \tilde{Z}_2^\varepsilon(ns)\right)\right|^{2p} \le C_p C_\gamma^p |t-s|^{\gamma p} \text{ uniformly in } |t-s| \in [0,1].$$

Choosing p such that $\gamma p > 1$, this bound gives the tightness [2, Theorem 13.5]. □

3.3 Tightness for $\tilde{Z}_1^\varepsilon$ and $\tilde{U}_1^\varepsilon$

Recall that $\varepsilon \in \{-1,1\}^{\mathbb{N}}$ is fixed. Let G be either $\tilde{Z}_1^\varepsilon$ or $\tilde{U}_1^\varepsilon$. To show the tightness, we will prove

$$\lim_{\delta\to0}\limsup_{n\to\infty}\mathbb{P}\left(\sup_{|t-s|\le\delta}|G(nt) - G(ns)| \ge \eta\sigma_n\right) = 0 \text{ for all } \eta > 0. \tag{3.4}$$

The tightness then follows from the corollary of Theorem 13.4 in [2]. To prove (3.4), we first show the following two lemmas.

Lemma 3.4. *Let G be either $\tilde{Z}_1^\varepsilon$ or $\tilde{U}_1^\varepsilon$. For all integer $p \ge 1$ and $\gamma \in (0,\alpha)$, there exits a constant $C_{p,\gamma} > 0$ such that for all $s,t \in [0,1]$, for all $n \ge 1$,*

$$\mathbb{E}|G(ns) - G(nt)|^{2p} \le C_{p,\gamma}\left(|t-s|^{\gamma p}\sigma_n^{2p} + |t-s|^\gamma\sigma_n^2\right). \tag{3.5}$$

Lemma 3.5. *Let G be either $\tilde{Z}_1^\varepsilon$ or $\tilde{U}_1^\varepsilon$. For all $t \le s \le t+\delta$,*

$$|G(t) - G(s)| \le N(t+\delta) - N(t) + \delta, \text{ almost surely,} \tag{3.6}$$

where N is the Poisson process in the definition of $\tilde{Z}_1^\varepsilon$ and $\tilde{U}_1^\varepsilon$.

A chaining argument is then applied to establish the tightness by proving the following.

Lemma 3.6. *If a process G satisfies* (3.5) *and* (3.6) *for a Poisson process N, then* (3.4) *holds.*

Proof of Lemma 3.4. We prove for $G = \tilde{U}_1^\varepsilon$. The case $G = \tilde{Z}_1^\varepsilon$ can be treated in a similar way and is omitted. In view of Lemma 3.1, it is sufficient to prove that for all $p \ge 1$ and all $0 \le s < t \le 1$,

$$\mathbb{E}|G(t) - G(s)|^{2p} \le C_p\Big(V(2(t-s))^p + V(2(t-s))\Big), \tag{3.7}$$

with the function V defined in (3.3). We prove (3.7) by induction on p. For $p = 1$, by independence of the N_k, we have

$$\mathbb{E}|G(t) - G(s)|^2 = \sum_{k\ge1}\mathrm{Var}\left(\mathbb{1}_{\{N_k(t)\text{ is odd}\}} - \mathbb{1}_{\{N_k(s)\text{ is odd}\}}\right)$$
$$\le \sum_{k\ge1}\mathbb{E}\left(\mathbb{1}_{\{N_k(t)\text{ is odd}\}} - \mathbb{1}_{\{N_k(s)\text{ is odd}\}}\right)^2$$
$$\le \sum_{k\ge1}\tilde{q}_k(t-s) = \frac{1}{2}V(2(t-s)).$$

Let $p \geq 2$ and assume that the property holds for $p-1$. We fix $0 < s < t$, and simplify the notations by setting

$$X_k := \mathbb{1}_{\{N_k(t) \text{ is odd}\}} - \tilde{q}_k(t) - \left(\mathbb{1}_{\{N_k(s) \text{ is odd}\}} - \tilde{q}_k(s)\right).$$

Note that $|X_k| \leq 2$ for all $k \geq 1$. Since $(X_k)_{k \geq 1}$ are centered and independent, it follows that

$$\begin{aligned}\mathbb{E}|G(t)-G(s)|^{2p} &= \sum_{k_1,\dots,k_p \geq 1} \mathbb{E}\left(X_{k_1}^2 \cdots X_{k_p}^2\right) \\ &\leq \sum_{\substack{k_1,\dots,k_p \geq 1 \\ k_1 \notin \{k_2,\dots,k_p\}}} \mathbb{E}\left(X_{k_1}^2\right) \mathbb{E}\left(X_{k_2}^2 \cdots X_{k_p}^2\right) + \sum_{\substack{k_1,\dots,k_p \geq 1 \\ k_1 \in \{k_2,\dots,k_p\}}} \mathbb{E}\left(X_{k_1}^2 \cdots X_{k_p}^2\right) \\ &\leq \left(\sum_{k_1 \geq 1} \mathbb{E}\left(X_{k_1}^2\right) + 4(p-1)\right) \sum_{k_2,\dots,k_p \geq 1} \mathbb{E}\left(X_{k_2}^2 \cdots X_{k_p}^2\right).\end{aligned}$$

By the induction hypothesis, we infer

$$\begin{aligned}\mathbb{E}|G(t)-G(s)|^{2p} &\leq \left(\frac{1}{2}V(2(t-s)) + 4(p-1)\right) C_{p-1}\Big(V(2(t-s))^{p-1} + V(2(t-s))\Big) \\ &\leq C'_p\Big(V(2(t-s))^p + V(2(t-s))^{p-1} + V(2(t-s))^2 + V(2(t-s))\Big),\end{aligned}$$

for a new positive constant C'_p depending only on p. We now deduce (3.7) using the fact that $V^\ell \leq V^p + V$ for all $1 < \ell < p$ and taking $C_p = 3C'_p$. □

Proof of Lemma 3.5. Let $t \leq s \leq t+\delta$. Recalling (3.1), we have

$$\begin{aligned}|\tilde{Z}_1^\varepsilon(s) - \tilde{Z}_1^\varepsilon(t)| &\leq \sum_{k \geq 1} \left|\mathbb{1}_{\{N_k(s) \neq 0\}} - \mathbb{1}_{\{N_k(t) \neq 0\}}\right| + \sum_{k \geq 1} |\tilde{p}_k(s) - \tilde{p}_k(t)| \\ &\leq \sum_{k \geq 1} \mathbb{1}_{\{N_k(s) - N_k(t) \neq 0\}} + \sum_{k \geq 1} \tilde{p}_k(s-t) \\ &\leq N(s) - N(t) + \mathbb{E}\left(N(s-t)\right) \\ &\leq N(t+\delta) - N(t) + \delta.\end{aligned}$$

Similarly, recalling (3.2),

$$\begin{aligned}|\tilde{U}_1^\varepsilon(s) - \tilde{U}_1^\varepsilon(t)| &\leq \sum_{k \geq 1} \left|\mathbb{1}_{\{N_k(s) \text{ is odd}\}} - \mathbb{1}_{\{N_k(t) \text{ is odd}\}}\right| + \sum_{k \geq 1} |\tilde{q}_k(s) - \tilde{q}_k(t)| \\ &\leq \sum_{k \geq 1} \mathbb{1}_{\{N_k(s) - N_k(t) \neq 0\}} + \sum_{k \geq 1} \tilde{q}_k(s-t) \\ &\leq N(t+\delta) - N(t) + \delta.\end{aligned}$$

□

Proof of Lemma 3.6. Let $\eta > 0$ be fixed. For $\delta \in (0,1)$ and $r := \left\lfloor \frac{1}{\delta} \right\rfloor + 1$, we set $t_i := i\delta$ for $i = 0, \dots, r-1$, and $t_r := 1$. By [2, Theorem 7.4], we have

$$\mathbb{P}\left(\sup_{|t-s| \leq \delta} |G(nt) - G(ns)| \geq 9\eta\sigma_n\right) \leq \sum_{i=1}^{r} \mathbb{P}\left(\sup_{t_{i-1} \leq s \leq t_i} |G(ns) - G(nt_{i-1})| \geq 3\eta\sigma_n\right). \tag{3.8}$$

The rest of the proof is based on a chaining argument. Fix $i \in \{1, \dots, r\}$. For all $k \geq 1$, we introduce the subdivision of rank k of the interval $[t_{i-1}, t_i]$:

$$x_{k,\ell} := t_{i-1} + \ell\frac{\delta}{2^k}, \text{ for } k \geq 1 \text{ and } \ell = 0, \dots, 2^k.$$

For $s \in [t_{i-1}, t_i]$ and $n \geq 1$, we define the chain $s_0 := t_{i-1} \leq s_1 \leq \ldots \leq s_{k_n} \leq s$, where for each k, s_k is the largest point among $(x_{k,\ell})_{\ell=0,\ldots,2^k}$ of rank k that is smaller than s, and where we choose

$$k_n := \left\lfloor \log_2 \left(2(e-1) \frac{n\delta}{\eta\sigma_n} \right) \right\rfloor + 1. \tag{3.9}$$

This choice of k_n will become clearer later. For $t_{i-1} \leq s \leq t_i$, we write

$$|G(ns) - G(nt_{i-1})| \leq \sum_{k=1}^{k_n} |G(ns_k) - G(ns_{k-1})| + |G(ns) - G(ns_{k_n})|, \tag{3.10}$$

and since we necessarily have $s_k = s_{k-1}$ or $s_k = s_{k-1} + \frac{\delta}{2^k}$, we infer that for all $k \geq 1$,

$$|G(ns_k) - G(ns_{k-1})| \leq \max_{\ell=1,\ldots,2^k} |G(nx_{k,\ell}) - G(nx_{k,\ell-1})|. \tag{3.11}$$

Now, by Lemma 3.5, we get

$$\begin{aligned} |G(ns) - G(ns_{k_n})| &\leq N(n(s_{k_n} + \delta 2^{-k_n})) - N(ns_{k_n}) + n\delta 2^{-k_n} \\ &\leq \max_{\ell=0,\ldots,2^{k_n}-1} \left(N(n(x_{k_n,\ell} + \delta 2^{-k_n})) - N(nx_{k_n,\ell}) \right) + n\delta 2^{-k_n}. \end{aligned} \tag{3.12}$$

Further, observe that our choice of k_n in (3.9) gives $n\delta 2^{-k_n} \leq \eta\sigma_n$. Using this last fact and the inequalities (3.10), (3.11), and (3.12), we infer

$$\begin{aligned} &\limsup_{n\to\infty} \mathbb{P}\left(\sup_{t_{i-1}\leq s\leq t_i} |G(ns) - G(nt_{i-1})| \geq 3\eta\sigma_n \right) \\ &\qquad \leq \limsup_{n\to\infty} \mathbb{P}\left(\sum_{k=1}^{k_n} \max_{\ell=1,\ldots,2^k} |G(nx_{k,\ell}) - G(nx_{k,\ell-1})| > \eta\sigma_n \right) \qquad (3.13) \\ &\qquad + \limsup_{n\to\infty} \mathbb{P}\left(\max_{\ell=0,\ldots,2^{k_n}-1} \left(N(n(x_{k_n,\ell} + \delta 2^{-k_n})) - N(nx_{k_n,\ell}) \right) > \eta\sigma_n \right). \qquad (3.14) \end{aligned}$$

For (3.14), using exponential Markov inequality and the fact that $\mathbb{E}(e^{N(x)}) = e^{x(e-1)}$, we infer

$$\begin{aligned} \mathbb{P}\left(\max_{\ell=0,\ldots,2^{k_n}-1} \left\{ N(n(x_{k_n,\ell} + \delta 2^{-k_n})) - N(nx_{k_n,\ell}) \right\} > \eta\sigma_n \right) &\leq 2^{k_n} \mathbb{P}\left(N(n\delta 2^{-k_n}) > \eta\sigma_n \right) \\ &\leq 2^{k_n} e^{n\delta 2^{-k_n}(e-1) - \eta\sigma_n}. \end{aligned}$$

Again, by the choice of k_n in (3.9), $2^{k_n} \leq 4(e-1)n\delta/(\eta\sigma_n)$ and $2^{-k_n} \leq \eta\sigma_n/(2(e-1)n\delta)$. Thus, the above inequality is bounded by $4(e-1)n\delta/(\eta\sigma_n)e^{-\frac{1}{2}\eta\sigma_n}$, which converges to 0 as $n \to \infty$. So, the term (3.14) vanishes and it remains to deal with (3.13). Let $\eta_k := \frac{\eta}{k(k+1)}$, $k \geq 1$, so that $\sum_{k\geq 1} \eta_k = \eta$. We have

$$\begin{aligned} &\mathbb{P}\left(\sum_{k=1}^{k_n} \max_{\ell=1,\ldots,2^k} |G(nx_{k,\ell}) - G(nx_{k,\ell-1})| > \eta\sigma_n \right) \\ &\qquad \leq \sum_{k=1}^{k_n} \mathbb{P}\left(\max_{\ell=1,\ldots,2^k} |G(nx_{k,\ell}) - G(nx_{k,\ell-1})| > \eta_k\sigma_n \right) \\ &\qquad \leq \sum_{k=1}^{k_n} \sum_{\ell=1}^{2^k} \mathbb{P}\left(|G(nx_{k,\ell}) - G(nx_{k,\ell-1})| > \eta_k\sigma_n \right). \end{aligned}$$

Now, fix $\gamma \in (0,\alpha)$ and let $p \geq 1$ be an integer such that $\gamma p > 1$. Using Markov inequality at order $2p$ and the $2p$-th moment bound (3.5) in Lemma 3.4, we get

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=1}^{k_n} \max_{\ell=1,\dots,2^k} |G(nx_{k,\ell}) - G(nx_{k,\ell-1})| > \eta\sigma_n\right)\\
&\qquad \leq \sum_{k=1}^{k_n}\sum_{\ell=1}^{2^k} \eta_k^{-2p} \frac{\mathbb{E}\,|G(nx_{k,\ell}) - G(nx_{k,\ell-1})|^{2p}}{\sigma_n^{2p}}\\
&\qquad \leq C_{p,\gamma}\sum_{k=1}^{k_n}\sum_{\ell=1}^{2^k} \eta_k^{-2p}\left(|x_{k,\ell} - x_{k,\ell-1}|^{\gamma p} + \frac{|x_{k,\ell} - x_{k,\ell-1}|^{\gamma}}{\sigma_n^{2(p-1)}}\right)\\
&\qquad \leq C_{p,\gamma}\delta^{\gamma p}\sum_{k=1}^{\infty}\eta_k^{-2p}2^{k(1-\gamma p)} + C_{p,\gamma}\delta^{\gamma} n^{\alpha(1-p)}L(n)^{1-p}\sum_{k=1}^{k_n}\eta_k^{-2p}2^{k(1-\gamma)}.
\end{aligned}$$

In the right-hand side, since $\gamma p > 1$, the series in the first term is converging and is independent of n. The sum in the second term is bounded, up to a multiplicative constant, by $2^{k_n(1-\gamma)}$ which is of order $n^{(1-\alpha/2)(1-\gamma)}$ (here and next line, up to a slowly varying function). Thus, the second term in the right-hand side is of order $n^{1-\alpha p+\alpha/2-\gamma+\gamma\alpha/2} \leq n^{1-\gamma p+(\alpha-\gamma)(1-p)}$ and vanishes as n goes to ∞, again because we have assumed $\gamma p > 1$. So for (3.13), we arrive at

$$\limsup_{n\to\infty}\mathbb{P}\left(\sum_{k=1}^{k_n}\max_{\ell=1,\dots,2^k}|G(nx_{k,\ell}) - G(nx_{k,\ell-1})| > \eta\sigma_n\right) \leq C\delta^{\gamma p}$$

for some constant C independent of δ and η. From (3.8), we conclude that

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{|t-s|\leq\delta}|G(nt) - G(ns)| \geq 9\eta\sigma_n\right) \leq C'\left(\left\lfloor\frac{1}{\delta}\right\rfloor + 1\right)\delta^{\gamma p}$$

which goes to 0 as $\delta \downarrow 0$. This yields (3.4). □

Remark 3.7. For the Poissonized model, we can establish similar weak convergence to the decompositions as in Theorems 2.1 and 2.2, by adapting the proofs at the end of Section 4. We omit this part.

4 De-Poissonization

In this section we prove our main theorems. Recall the decompositions

$$Z^\varepsilon = Z_1^\varepsilon + Z_2^\varepsilon \quad \text{and} \quad U^\varepsilon = U_1^\varepsilon + U_2^\varepsilon,$$

and

$$\tilde{Z}^\varepsilon = \tilde{Z}_1^\varepsilon + \tilde{Z}_2^\varepsilon \quad \text{and} \quad \tilde{U}^\varepsilon = \tilde{U}_1^\varepsilon + \tilde{U}_2^\varepsilon.$$

Note that G^ε and $\tilde{G}^\varepsilon$, for G being Z_1, Z_2, U_1, U_2 respectively, are coupled in the sense that they are defined on the same probability space as functionals of the same ε and $(Y_n)_{n\geq 1}$. We have already established weak convergence results for $\tilde{Z}_1^\varepsilon, \tilde{Z}_2^\varepsilon, \tilde{U}_1^\varepsilon, \tilde{U}_2^\varepsilon$. The de-Poissonization step thus consists of controlling the distance between G^ε and $\tilde{G}^\varepsilon$. We first prove the easier part.

4.1 The processes Z_2^ε and U_2^ε

Theorem 4.1. *For a Rademacher sequence ε,*

$$\left(\frac{Z_2^\varepsilon(\lfloor nt\rfloor)}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{Z}_2(t))_{t\in[0,1]} \quad \textit{and} \quad \left(\frac{U_2^\varepsilon(\lfloor nt\rfloor)}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{U}_2(t))_{t\in[0,1]},$$

in $D([0,1])$, where $\mathbb{Z}_2$ and $\mathbb{U}_2$ are as in Theorems 2.1 and 2.2.

Proof. Thanks to the coupling, it suffices to show for all $\varepsilon \in \{-1,1\}^{\mathbb{N}}$ fixed,

$$\lim_{n\to\infty} \sup_{t\in[0,1]} \frac{|\widetilde{G}^\varepsilon(nt) - G^\varepsilon(\lfloor nt\rfloor)|}{\sigma_n} = 0$$

in probability, with G being Z_2, U_2 respectively. We actually prove the above convergence in the almost sure sense. Observe that for all $\varepsilon \in \{-1,1\}^{\mathbb{N}}$,

$$|\widetilde{Z}_2^\varepsilon(nt) - Z_2^\varepsilon(\lfloor nt\rfloor)| \le \sum_{k\ge1} |\tilde{p}_k(nt) - p_k(\lfloor nt\rfloor)|,$$
$$|\widetilde{U}_2^\varepsilon(nt) - U_2^\varepsilon(\lfloor nt\rfloor)| \le \sum_{k\ge1} |\tilde{q}_k(nt) - q_k(\lfloor nt\rfloor)|.$$

Thus, the proof is completed once the following lemma is proved. □

Lemma 4.2. *The following limits hold:*

$$\lim_{n\to\infty} \frac{1}{\sigma_n} \sup_{t\in[0,1]} \sum_{k\ge1} |\tilde{p}_k(nt) - p_k(\lfloor nt\rfloor)| = 0 \tag{4.1}$$

and

$$\lim_{n\to\infty} \frac{1}{\sigma_n} \sup_{t\in[0,1]} \sum_{k\ge1} |\tilde{q}_k(nt) - q_k(\lfloor nt\rfloor)| = 0. \tag{4.2}$$

Proof. By triangular inequality, for all $n \ge 1$, $t \ge 0$,

$$\sum_{k\ge1} |\tilde{p}_k(nt) - p_k(\lfloor nt\rfloor)| \le \sum_{k\ge1} |\tilde{p}_k(\lfloor nt\rfloor) - \tilde{p}_k(nt)| + \sum_{k\ge1} |\tilde{p}_k(\lfloor nt\rfloor) - p_k(\lfloor nt\rfloor)|.$$

First, note that for all $k \ge 1$,

$$|\tilde{p}_k(\lfloor nt\rfloor) - \tilde{p}_k(nt)| \le \tilde{p}_k(\lfloor nt\rfloor + 1) - \tilde{p}_k(\lfloor nt\rfloor) = e^{-p_k\lfloor nt\rfloor}(1 - e^{-p_k}),$$

and thus,

$$\sum_{k\ge1} |\tilde{p}_k(\lfloor nt\rfloor) - \tilde{p}_k(nt)| \le \sum_{k\ge1} p_k = 1.$$

Further, if $\lfloor nt\rfloor \ge 1$, using that $e^{-my} - (1-y)^m \le \frac{1}{m}(1 - e^{-my})$ for all $0 \le y \le 1$ and $m \in \mathbb{N}$, we have

$$\begin{aligned}\sum_{k\ge1} |\tilde{p}_k(\lfloor nt\rfloor) - p_k(\lfloor nt\rfloor)| &= \sum_{k\ge1} \left(e^{-p_k\lfloor nt\rfloor} - (1-p_k)^{\lfloor nt\rfloor}\right) \\ &\le \frac{1}{\lfloor nt\rfloor} \sum_{k\ge1} (1 - e^{-p_k\lfloor nt\rfloor}) = \frac{V(\lfloor nt\rfloor)}{\lfloor nt\rfloor},\end{aligned}$$

which is bounded (since $V(n)/n \to 0$ as $n \to \infty$). We thus deduce (4.1). The proof for (4.2) is similar and omitted. □

4.2 The processes Z_1^ε and U_1^ε

In this section we prove Theorem 2.3. The coupling of $Z_1^\varepsilon, \tilde{Z}_1^\varepsilon$ and $U_1^\varepsilon, \tilde{U}_1^\varepsilon$ respectively takes a little more effort to control.

Proof of Theorem 2.3. Let N be the Poisson process introduced in Section 3 and denote by τ_i the i-th arrival time of N, $i \geq 1$, namely $\tau_i := \inf\{t > 0 \mid N(t) = i\}$. We introduce the random changes of time $\lambda_n : [0, \infty) \to [0, \infty)$, $n \geq 1$, given by

$$\lambda_n(t) := \frac{\tau_{\lfloor nt \rfloor}}{n}, \quad t \geq 0.$$

By constructions, we have

$$Z^\varepsilon(\lfloor nt \rfloor) = \tilde{Z}^\varepsilon(n\lambda_n(t)) \quad \text{and} \quad \tilde{U}^\varepsilon(\lfloor nt \rfloor) = U^\varepsilon(n\lambda_n(t)), \text{ almost surely.}$$

These identities do not hold for the process Z_1^ε or U_1^ε but we can still couple $Z_1^\varepsilon, \tilde{Z}_1^\varepsilon$ and $U_1^\varepsilon, \tilde{U}_1^\varepsilon$ via

$$Z_1^\varepsilon(\lfloor nt \rfloor) = \tilde{Z}_1^\varepsilon(n\lambda_n(t)) + \sum_{k \geq 1} \varepsilon_k(\tilde{p}_k(n\lambda_n(t)) - p_k(\lfloor nt \rfloor)) \tag{4.3}$$

$$U_1^\varepsilon(\lfloor nt \rfloor) = \tilde{U}_1^\varepsilon(n\lambda_n(t)) + \sum_{k \geq 1} \varepsilon_k(\tilde{q}_k(n\lambda_n(t)) - q_k(\lfloor nt \rfloor)). \tag{4.4}$$

The proof is now decomposed into two lemmas treating separately the two terms in the right-hand side of the preceding identities.

Lemma 4.3. *We have*

$$\left(\frac{\tilde{Z}_1^\varepsilon(n\lambda_n(t))}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{Z}_1(t))_{t\in[0,1]} \quad \textit{and} \quad \left(\frac{\tilde{U}_1^\varepsilon(n\lambda_n(t))}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{U}_1(t))_{t\in[0,1]}$$

in $D([0,1])$.

Proof. We only prove the first convergence. The proof of the second is the same by replacing $(\tilde{Z}_1^\varepsilon, \mathbb{Z}_1)$ by $(\tilde{U}_1^\varepsilon, \mathbb{U}_1)$. For $t \geq 0$, by the law of large numbers, $\lambda_n(t) \to t$ almost surely as $n \to \infty$. Since the λ_n are nondecreasing, almost surely the convergence holds for all $t \geq 0$, and by Pólya's extension of Dini's theorem (see [23, Problem 127]) the convergence is uniform for t in a compact interval. That is

$$\lim_{n\to\infty} \sup_{t\in[0,1]} |\lambda_n(t) - t| = 0 \text{ almost surely,}$$

and λ_n converges almost surely to the identity function $\mathbb{I}$ in $D([0,1])$.

We want to apply the random change of time lemma from Billingsley [2, p. 151]. However, λ_n is not a good candidate as it is not bounded between $[0,1]$. Instead, we introduce

$$\lambda_n^*(t) := \min\left(\lambda_n(t), 1\right), \quad t \geq 0.$$

Observe that by monotonicity,

$$\sup_{t\in[0,1]} |\lambda_n^*(t) - t| \leq \sup_{t\in[0,1]} |\lambda_n(t) - t|.$$

Thus, λ_n^* converges almost surely to $\mathbb{I}$ in $D([0,1])$. By Slutsky's lemma and Proposition 3.2, we also have

$$\left(\left(\frac{\tilde{Z}_1^\varepsilon(nt)}{\sigma_n}\right)_{t\in[0,1]}, (\lambda_n^*(t))_{t\in[0,1]}\right) \Rightarrow \left((\mathbb{Z}_1(t))_{t\in[0,1]}, \mathbb{I}\right) \tag{4.5}$$

in $D([0,1]) \times D([0,1])$. Furthermore, since λ_n^* is non-decreasing and bounded in $[0,1]$, thus by random change of time lemma we obtain

$$\left(\frac{\tilde{Z}_1^\varepsilon(n\lambda_n^*(t))}{\sigma_n}\right)_{t\in[0,1]} \Rightarrow (\mathbb{Z}_1(t))_{t\in[0,1]} \tag{4.6}$$

in $D([0,1])$. To obtain the desired result we need to replace λ_n^* by λ_n. However, by definition, for all $\eta \in (0,1)$ fixed,

$$\mathbb{P}(\lambda_n^* \neq \lambda_n \text{ on } [0,1-\eta]) \leq \mathbb{P}\left(\tau_{\lfloor n(1-\eta)\rfloor} \geq n\right) \to 0 \text{ as } n \to \infty.$$

It then follows that, restricting the convergence of (4.6) in $D([0,1-\eta])$,

$$\left(\frac{\tilde{Z}_1^\varepsilon(n\lambda_n(t))}{\sigma_n}\right)_{t\in[0,1-\eta]} \Rightarrow (\mathbb{Z}_1(t))_{t\in[0,1-\eta]}$$

in $D([0,1-\eta])$. This is strictly weaker than the convergence in $D([0,1])$ that we are looking for. However, looking back we see an easy fix as follows. If one starts in (4.5) with weak convergence for $\tilde{Z}_1^\varepsilon$ and λ_n^* (modified accordingly) as processes indexed by a slightly larger time interval, say in $D([0,1/(1-\eta)])$ for any $\eta \in (0,1)$ fixed, the desired result then follows. □

In view of Lemma 4.2, the following lemma will be sufficient to conclude.

Lemma 4.4. *The following limits hold:*

$$\lim_{n\to\infty} \frac{1}{\sigma_n} \sup_{t\in[0,1]} \sum_{k\geq 1} |\tilde{p}_k(nt) - \tilde{p}_k(n\lambda_n(t))| = 0 \textit{ in probability}$$

and

$$\lim_{n\to\infty} \frac{1}{\sigma_n} \sup_{t\in[0,1]} \sum_{k\geq 1} |\tilde{q}_k(nt) - \tilde{q}_k(n\lambda_n(t))| = 0 \textit{ in probability.} \tag{4.7}$$

Proof. We only prove the second limit. The first one can be proved in a similar way and is omitted. We first introduce

$$\Lambda_n(t) := n^{\frac{1}{2}}(\lambda_n(t) - t) = n^{-\frac{1}{2}}(\tau_{\lfloor nt\rfloor} - nt).$$

Since τ_n is the sum of i.i.d. random variables with exponential distribution of rate 1, and since $n^{-\frac{1}{2}}(nt - \lfloor nt\rfloor)$ converges to 0 uniformly in t, by Donsker's theorem and Slutsky's lemma, we have

$$(\Lambda_n(t))_{t\in[0,1]} \Rightarrow (\mathbb{B}(t))_{t\in[0,1]} \text{ in } D([0,1]),$$

where $\mathbb{B}$ is a standard Brownian motion. By the continuous mapping theorem, the sequence $\sup_{t\in[0,1]} |\Lambda_n(t)|$ weakly converges to $\sup_{t\in[0,1]} |\mathbb{B}(t)|$, as $n \to \infty$. In particular, $(\sup_{t\in[0,1]} |\Lambda_n(t)|)_{n\geq 1}$ is tight. So, for any $\eta > 0$, there exits $K_\eta > 0$ such that for n large enough,

$$\mathbb{P}\left(\sup_{t\in[0,1]} |\Lambda_n(t)| > K_\eta\right) \leq \eta. \tag{4.8}$$

Now, choose $\beta \in (0,1/2)$ and consider

$$A_n := \sup_{t\in[0,n^{-\beta}]} \sum_{k\geq 1} |\tilde{q}_k(nt) - \tilde{q}_k(n\lambda_n(t))| \quad \text{and} \quad B_n := \sup_{t\in[n^{-\beta},1]} \sum_{k\geq 1} |\tilde{q}_k(nt) - \tilde{q}_k(n\lambda_n(t))|.$$

Concerning A_n, using the bound in (3.2), we have

$$A_n \leq \sup_{t\in[0,n^{-\beta}]} \sum_{k\geq 1} \tilde{q}_k(n|\lambda_n(t)-t|) = \sup_{t\in[0,1]} \sum_{k\geq 1} \tilde{q}_k(n|\lambda_n(n^{-\beta}t)-n^{-\beta}t|).$$

We can write

$$\lambda_n(n^{-\beta}t) - n^{-\beta}t = \frac{\Lambda_{n^{1-\beta}}(t)}{n^{\frac{1+\beta}{2}}}.$$

For any $\eta > 0$, using (4.8), by monotonicity of $\tilde{q}_k(\cdot)$, we infer that for n large enough

$$\mathbb{P}\left(A_n \leq \sum_{k\geq 1} \tilde{q}_k\left(n\cdot n^{-\frac{1+\beta}{2}}K_\eta\right)\right) > 1-\eta.$$

But

$$\begin{aligned}\frac{1}{\sigma_n}\sum_{k\geq 1}\tilde{q}_k\left(n^{1-(1+\beta)/2}K_\eta\right) &= \frac{1}{2\sigma_n}V\left(2n^{(1-\beta)/2}K_\eta\right) \\ &\sim \Gamma(1-\alpha)2^{\alpha-1}K_\eta^\alpha n^{-\beta\alpha/2}\frac{L(n^{(1-\beta)/2})}{L(n)^{1/2}} \longrightarrow 0 \text{ as } n\to\infty.\end{aligned}$$

Thus, A_n/σ_n converges to 0 in probability as n goes to ∞.
Concerning B_n, using the identity (3.2), we can write

$$\begin{aligned}B_n &= \sup_{t\in[n^{-\beta},1]}\sum_{k\geq 1}\left(1-2\tilde{q}_k(n\min(\lambda_n(t),t))\right)\tilde{q}_k(n|\lambda_n(t)-t|) \\ &= \sup_{t\in[n^{-\beta},1]}\sum_{k\geq 1} e^{-2p_k n\min(\lambda_n(t),t)}\tilde{q}_k(n|\lambda_n(t)-t|).\end{aligned}$$

Now, for $t\in[n^{-\beta},1]$, observe that if for some $K>0$, $|\Lambda_n(t)|\leq K$ and $n^{\frac{1}{2}-\beta} > 2K$, then

$$\lambda_n(t) = t + n^{-1/2}\Lambda_n(t) \geq t - n^{-1/2}|\Lambda_n(t)| \geq t - \frac{n^{-\beta}}{2} \geq \frac{t}{2},$$

and thus $\min(\lambda_n(t),t)\geq \frac{t}{2}$. Let $\eta>0$ and K_η be as in (4.8). Assume n is large enough so that (4.8) holds and $n^{\frac{1}{2}-\beta} > 2K_\eta$ (which is possible since we have chosen $\beta\in(0,1/2)$). By the preceding observation and by monotonicity of $\tilde{q}_k(\cdot)$, we infer

$$\mathbb{P}\left(B_n \leq \sup_{t\in[n^{-\beta},1]}\sum_{k\geq 1} e^{-p_k nt}\tilde{q}_k\left(n\cdot n^{-\frac{1}{2}}K_\eta\right)\right) > 1-\eta.$$

Now, using $1-e^{-x}\leq x$ and then $xe^{-x}\leq 1-e^{-x}$ for $x>0$, we get

$$\begin{aligned}\sup_{t\in[n^{-\beta},1]}\sum_{k\geq 1} e^{-p_k nt}\tilde{q}_k\left(n^{1-\frac{1}{2}}K_\eta\right) &= \sum_{k\geq 1} e^{-p_k n^{1-\beta}}\frac{1}{2}\left(1-e^{-2p_k n^{\frac{1}{2}}K_\eta}\right) \\ &\leq \sum_{k\geq 1} e^{-p_k n^{1-\beta}} p_k n^{\frac{1}{2}}K_\eta \\ &\leq \sum_{k\geq 1}\left(1-e^{-p_k n^{1-\beta}}\right) n^{-\frac{1}{2}+\beta}K_\eta = n^{\beta-1/2}V(n^{1-\beta})K_\eta.\end{aligned}$$

Thus,

$$\begin{aligned}\frac{1}{\sigma_n}\sup_{t\in[n^{-\beta},1]}\sum_{k\geq 1} e^{-p_k nt}\tilde{q}_k\left(n^{1-\frac{1}{2}}K_\eta\right) &\leq \frac{n^{\beta-1/2}V(n^{1-\beta})}{\sigma_n}K_\eta \\ &\sim \Gamma(1-\alpha)K_\eta n^{(\beta-\frac{1}{2})(1-\alpha)}\frac{L(n^{(1-\beta)})}{L(n)^{1/2}} \longrightarrow 0 \text{ as } n\to\infty,\end{aligned}$$

since $\beta \in (0,1/2)$. Thus B_n/σ_n converges to 0 in probability as n goes to ∞. We have thus proved (4.7). □

To sum up, the desired results now follow from (4.3) and (4.4), Lemmas 4.2, 4.3 and 4.4, and Slutsky's lemma. □

4.3 The trivariate processes

Finally we conclude by establishing the main theorems.

Proof of Theorems 2.1 and 2.2. We prove Theorem 2.1. The proof for Theorem 2.2 is the same. We denote by $\mathcal{E}$ the σ-field generated by the $(\varepsilon_k)_{k\geq 1}$ which is then independent of $(Y_n)_{n\geq 1}$. Note that the process Z_2^ε is $\mathcal{E}$-measurable. For any continuous and bounded function f and g from $D([0,1])$ to $\mathbb{R}$, we have

$$\begin{aligned}
&\left|\mathbb{E}\left(f\left(\frac{Z_1^\varepsilon(\lfloor n\cdot\rfloor)}{\sigma_n}\right)g\left(\frac{Z_2^\varepsilon(\lfloor n\cdot\rfloor)}{\sigma_n}\right)\right)-\mathbb{E}f(\mathbb{Z}_1)\mathbb{E}g(\mathbb{Z}_2)\right|\\
&\quad=\left|\mathbb{E}\left[\mathbb{E}\left(f\left(\frac{Z_1^\varepsilon(\lfloor n\cdot\rfloor)}{\sigma_n}\right)\,\middle|\,\mathcal{E}\right)g\left(\frac{Z_2^\varepsilon(\lfloor n\cdot\rfloor)}{\sigma_n}\right)\right]-\mathbb{E}f(\mathbb{Z}_1)\mathbb{E}g(\mathbb{Z}_2)\right|\\
&\quad\leq\mathbb{E}\left|\mathbb{E}\left(f\left(\frac{Z_1^\varepsilon(\lfloor n\cdot\rfloor)}{\sigma_n}\right)\,\middle|\,\mathcal{E}\right)-\mathbb{E}f(\mathbb{Z}_1)\right|\cdot\|g\|_\infty+\left|\mathbb{E}g\left(\frac{Z_2^\varepsilon(\lfloor n\cdot\rfloor)}{\sigma_n}\right)-\mathbb{E}g(\mathbb{Z}_2)\right|\cdot\|f\|_\infty.
\end{aligned}$$

The first term goes to 0 as $n\to\infty$ thanks to Theorem 2.3 and the dominated convergence theorem. The second one goes to 0 as $n\to\infty$ thanks to Theorem 4.1. By [28, Corollary 1.4.5] we deduce that

$$\frac{1}{\sigma_n}\left(Z_1^\varepsilon(\lfloor nt\rfloor),Z_2^\varepsilon(\lfloor nt\rfloor)\right)_{t\in[0,1]}\Rightarrow\left(\mathbb{Z}_1(t),\mathbb{Z}_2(t)\right)_{t\in[0,1]},$$

in $D([0,1])^2$ where $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are independent. The rest of the theorem follows from the identity $Z^\varepsilon = Z_1^\varepsilon + Z_2^\varepsilon$. □

References

[1] Raghu R. Bahadur, *On the number of distinct values in a large sample from an infinite discrete distribution*, Proc. Nat. Inst. Sci. India Part A **26** (1960), no. supplement II, 67–75. MR-0137256

[2] Patrick Billingsley, *Convergence of probability measures*, second ed., Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Inc., New York, 1999, A Wiley-Interscience Publication. MR-1700749

[3] Nicholas H. Bingham, Charles M. Goldie, and Józef L. Teugels, *Regular variation*, Encyclopedia of Mathematics and its Applications, vol. 27, Cambridge University Press, Cambridge, 1987. MR-0898871

[4] Tomasz Bojdecki, Luis G. Gorostiza, and Anna Talarczyk, *Sub-fractional Brownian motion and its relation to occupation times*, Statist. Probab. Lett. **69** (2004), no. 4, 405–419. MR-2091760

[5] Tomasz Bojdecki and Anna Talarczyk, *Particle picture interpretation of some Gaussian processes related to fractional Brownian motion*, Stochastic Process. Appl. **122** (2012), no. 5, 2134–2154. MR-2921975

[6] J. Bunge and M. Fitzpatrick, *Estimating the number of species: a review*, J. Am. Stat. Ass. **88** (1993), no. 421, 364–373.

[7] Ju. A. Davydov, *The invariance principle for stationary processes*, Teor. Verojatnost. i Primenen. **15** (1970), 498–509. MR-0283872

[8] Kacha Dzhaparidze and Harry van Zanten, *A series expansion of fractional Brownian motion*, Probab. Theory Related Fields **130** (2004), no. 1, 39–55. MR-2092872

[9] Paul Embrechts and Makoto Maejima, *Selfsimilar processes*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, 2002. MR-1920153

[10] Nathanaël Enriquez, *A simple construction of the fractional Brownian motion*, Stochastic Process. Appl. **109** (2004), no. 2, 203–223. MR-2031768

[11] Alexander Gnedin, Ben Hansen, and Jim Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws*, Probab. Surv. **4** (2007), 146–171. MR-2318403

[12] Alan Hammond and Scott Sheffield, *Power law Pólya's urn and fractional Brownian motion*, Probab. Theory Related Fields **157** (2013), no. 3–4, 691–719. MR-3129801

[13] Christian Houdré and José Villa, *An example of infinite dimensional quasi-helix*, Stochastic models (Mexico City, 2002), Contemp. Math., vol. 336, Amer. Math. Soc., Providence, RI, 2003, pp. 195–201. MR-2037165

[14] Samuel Karlin, *Central limit theorems for certain infinite urn schemes*, J. Math. Mech. **17** (1967), 373–401. MR-0216548

[15] Claudia Klüppelberg and Christoph Kühn, *Fractional Brownian motion as a weak limit of Poisson shot noise processes—with applications to finance*, Stochastic Process. Appl. **113** (2004), no. 2, 333–351. MR-2087964

[16] Andreĭ N. Kolmogorov, *Wienersche Spiralen und einige andere interessante Kurven im Hilbertschen Raum*, C. R. (Doklady) Acad. Sci. URSS (N.S.) **26** (1940), 115–118. MR-0003441

[17] Pedro Lei and David Nualart, *A decomposition of the bifractional Brownian motion and some applications*, Statist. Probab. Lett. **79** (2009), no. 5, 619–624. MR-2499385

[18] Benoit B. Mandelbrot and John W. Van Ness, *Fractional Brownian motions, fractional noises and applications*, SIAM Rev. **10** (1968), 422–437. MR-0242239

[19] Thomas Mikosch and Gennady Samorodnitsky, *Scaling limits for cumulative input processes*, Math. Oper. Res. **32** (2007), no. 4, 890–918. MR-2363203

[20] Magda Peligrad and Sunder Sethuraman, *On fractional Brownian motion limits in one dimensional nearest-neighbor symmetric simple exclusion*, ALEA Lat. Am. J. Probab. Math. Stat. **4** (2008), 245–255. MR-2448774

[21] Vladas Pipiras and Murad S. Taqqu, *Long-range dependence and self-similarity*, Cambridge University Press, Forthcoming in 2016.

[22] Jim Pitman, *Combinatorial stochastic processes*, Lecture Notes in Mathematics, vol. 1875, Springer-Verlag, Berlin, 2006, Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard. MR-2245368

[23] George Pólya and Gábor Szegő, *Problems and theorems in analysis. Vol. I: Series, integral calculus, theory of functions*, Springer-Verlag, New York-Berlin, 1972. MR-0344042

[24] Juan Ruiz de Chávez and Constantin Tudor, *A decomposition of sub-fractional Brownian motion*, Math. Rep. (Bucur.) **11(61)** (2009), no. 1, 67–74. MR-2506510

[25] Gennady Samorodnitsky, *Long range dependence*, Found. Trends Stoch. Syst. **1** (2006), no. 3, 163–257. MR-2379935

[26] Frank Spitzer, *Principles of random walk*, The University Series in Higher Mathematics, D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto-London, 1964. MR-0171290

[27] Murad S. Taqqu, *Weak convergence to fractional Brownian motion and to the Rosenblatt process*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **31** (1975), 287–302. MR-0400329

[28] Aad W. van der Vaart and Jon A. Wellner, *Weak convergence and empirical processes*, Springer Series in Statistics, Springer-Verlag, New York, 1996, With applications to statistics. MR-1385671

Acknowledgments. The authors would like to thank David Nualart and Gennady Samorodnitsky for helpful discussions. The first author would like to thank the hospitality and financial support from Taft Research Center and Department of Mathematical Sciences at University of Cincinnati, for his visit in May and June 2015, when most of the results were obtained. The first author's research was partially supported by the

Région Centre project MADACA. The second author would like to thank the invitation and hospitality of Laboratoire de Mathématiques et Physique Théorique, UMR-CNRS 7350, Tours, France, for his visit from April to July in 2014, when the project was initiated. The second author's research was partially supported by NSA grant H98230-14-1-0318.