BINARY SEARCH TREES OF PERMUTON SAMPLES

BENOÎT CORSINI, VICTOR DUBACH, AND VALENTIN FÉRAY

ABSTRACT. Binary search trees (BST) are a popular type of data structure when dealing with ordered data. Indeed, they enable one to access and modify data efficiently, with their height corresponding to the worst retrieval time. From a probabilistic point of view, binary search trees associated with data arriving in a uniform random order are well understood, but less is known when the input is a non-uniform random permutation.

We consider here the case where the input comes from i.i.d. random points in the plane with law μ, a model which we refer to as a *permuton sample*. Our results show that the asymptotic proportion of nodes in each subtree depends on the behavior of the measure μ at its left boundary, while the height of the BST has a universal asymptotic behavior for a large family of measures μ. Our approach involves a mix of combinatorial and probabilistic tools, namely combinatorial properties of binary search trees, coupling arguments, and deviation estimates.

1. INTRODUCTION

1.1. **Context and informal description of our results.** A *binary search tree* (BST) is a rooted binary tree where nodes carry labels (which are real numbers) and where, for each vertex v, all labels of vertices in the left-subtree (resp. right-subtree) attached to v are smaller (resp. bigger) than the label of v. Binary search trees are a popular type of data structure for storing ordered data. One key feature is that the worst-case complexity of basic operations (lookup, addition or removal of data) is proportional to the height of the tree.

Given a BST $\mathcal{T}$ and a real number x distinct from the labels of $\mathcal{T}$, there is a unique way to insert x into $\mathcal{T}$, i.e. there is a unique BST $\mathcal{T}^{+x}$ obtained from $\mathcal{T}$ by adding a new node with label x. Iterating this operation starting from the empty tree and a sequence $y = (y_1, \ldots, y_n)$ of distinct values, we get a BST $\mathcal{T}\langle y \rangle$ with n nodes. An example of the sequence of trees obtained from $y = (2, 4, 1, 6, 3, 5)$ can be found in Figure 1. The shape of $\mathcal{T}\langle y \rangle$ (i.e. the underlying binary tree without node labels) depends only on the relative order of the numbers $y_1, \ldots, y_n$, and not on their actual value. We can thus assume without loss of generality that the sequence y is a permutation σ of the integers from 1 to n, and write $\mathcal{T}\langle \sigma \rangle = \mathcal{T}\langle \sigma_1, \ldots, \sigma_n \rangle$ in this case.

In the worst case, the tree $\mathcal{T}\langle \sigma \rangle$ has height $n - 1$ and further operations such as lookup, addition or removal of data will have a linear complexity, which is far from optimal. However it has been proven by Devroye [Dev86] that, if σ is uniformly distributed in the symmetric group S_n, then the height $h(\mathcal{T}\langle \sigma \rangle)$ is asymptotically equivalent to $c^* \log(n)$ for some constant c^*, yielding a much better complexity for later operations. Assuming that σ is uniformly distributed means that the data used to construct our BST arrived in a completely random order, which is in general unrealistic. It seems therefore natural to study BSTs associated with non-uniform random permutations, and in particular to see how Devroye's result is modified when changing the distribution of σ.

A first step in this direction has been performed in the papers [ABC21, Cor23], where the BSTs associated with random Mallows and record-biased permutations are studied, showing interesting phase transition phenomena. In the current paper, we will consider some geometric models of random permutations, sampled via i.i.d. random points in the plane with some common distribution μ. These models will be referred to here as *permuton samples*, and denoted by σ_μ^n; they appear naturally in a recently developed theory of limiting objects for large permutations, called permutons [HKM+13]. The goal of studying such models is twofold. First, it is a much larger but still tractable family of models than those considered before (permuton samples are indexed by probability measures on the square,
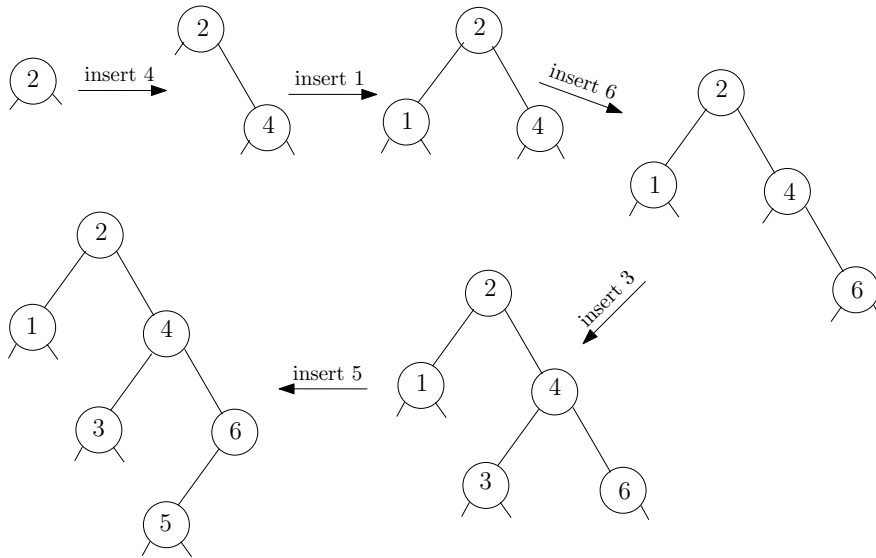
FIGURE 1. Iterative construction of the BST associated with the sequence $y = (2, 4, 1, 6, 3, 5)$. Let us detail the step where 3 is inserted. Since 3 is bigger than the root label (here 2), it should be added in the right-subtree attached to the root. We then compare 3 to the label of the root of that subtree, which is 4 in our example. Since 3 is smaller than 4, it should be added in the left subtree attached to 4. This subtree is empty at this stage, so we simply attach 3 to the left of 4.

while Mallows and record-biased permutations are one-parameter families of models). Second, since permutons describe the "large-scale shape" of permutations, it enlightens the connection between this "large-scale shape" and the associated BST.

Our first result (Theorem 1.1) shows that for a large family of permuton samples, the asymptotic behavior of the BST height is the same as the one found by Devroye for uniform random permutations, namely $h\left(\mathcal{T}\langle\sigma_\mu^n\rangle\right)$ is asymptotically equivalent to $c^* \log(n)$. We also consider the repartition of nodes in various branches of the BST, using the formalism of subtree size convergence recently introduced by Grübel in [Grü23]. In this setting, Theorem 1.2 below proves the convergence of the BST associated with permuton samples, under some mild assumption, where the limit object depends on the permuton only through its "derivative" at the left edge of the unit square $[0, 1]^2$ (the *derivative* of a permuton at $\{0\} \times [0, 1]$ does not make sense in general, but the mild assumption in the theorem precisely postulates its existence).

In the remaining part of the introduction, we present the model of permuton samples (Section 1.2), state our results precisely (Sections 1.3 and 1.4) and give an overview of the proofs (Section 1.5).

1.2. **Our model: binary search trees of permuton samples.** There is a natural way to map a (generic) finite set of points $\mathcal{P} \subset \mathbb{R}^2$ to a permutation $\sigma\langle\mathcal{P}\rangle$ and a binary search tree $\mathcal{T}\langle\mathcal{P}\rangle$, which we describe now. Let $\mathcal{P} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a set of points in $\mathbb{R}^2$ with distinct x- and distinct y-coordinates, and let $\{(x_{(1)}, y_{(1)}), \ldots, (x_{(n)}, y_{(n)})\}$ be its reordering such that $x_{(1)} < \ldots < x_{(n)}$. Then there exists a unique permutation $\sigma = \sigma\langle\mathcal{P}\rangle$ such that $(y_{(1)}, \ldots, y_{(n)})$ and $(\sigma_1, \ldots, \sigma_n)$ are in the same relative order. We let $\mathcal{T}\langle\mathcal{P}\rangle := \mathcal{T}\langle y_{(1)}, \ldots, y_{(n)}\rangle$ and note that the trees $\mathcal{T}\langle\mathcal{P}\rangle = \mathcal{T}\langle y_{(1)}, \ldots, y_{(n)}\rangle$ and $\mathcal{T}\langle\sigma\langle\mathcal{P}\rangle\rangle = \mathcal{T}\langle\sigma_1, \ldots, \sigma_n\rangle$ have the same shape since the two sequences have the same relative order. They do, however, have different sets of labels: $\{y_1, \ldots, y_n\}$ for $\mathcal{T}\langle\mathcal{P}\rangle$ and $\{1, \ldots, n\}$ for $\mathcal{T}\langle\sigma\langle\mathcal{P}\rangle\rangle$. These constructions are illustrated in Figure 2. The shape of $\mathcal{T}\langle\mathcal{P}\rangle$ is indeed the same as that of $\mathcal{T}\langle\sigma\langle\mathcal{P}\rangle\rangle$ (see the last image in Figure 1).
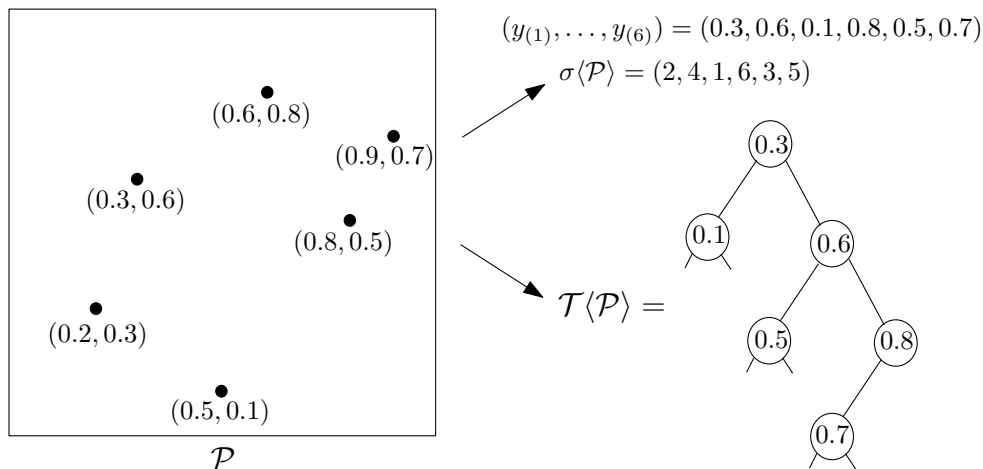
FIGURE 2. A set of points in $\mathbb{R}^2$ and its associated permutation and binary search tree.

Now consider a probability measure μ on $\mathbb{R}^2$ and take a set $\mathcal{P}_\mu^n$ of n i.i.d. points in $\mathbb{R}^2$ with distribution μ. In order to make sure that the associated permutation and binary search tree are always well-defined, we need the coordinates of the points to be all distinct. To this extent, we assume for the rest of this work that the projections of μ on both axes have no atom. Moreover, since the permutation and the shape of the tree only depend on the relative positions of the points, without loss of generality we can re-scale μ so that its support is in $[0,1]^2$ and so that, for any Lebesgue-measurable subset A of $[0,1]$:

$$\mu(A \times [0,1]) = \mu([0,1] \times A) = \lambda(A) \tag{1}$$

where λ is the uniform measure on $[0,1]$ (see [BDMW23, Remark 1.2] for details). Such measures μ with uniform projections on the axes are called *permutons*. Permutons are natural limit objects for large permutations, see e.g. [HKM+13, BBF+20]. The associated model of random permutations $\sigma\langle\mathcal{P}_\mu^n\rangle$ will then simply be denoted by σ_μ^n. This is a broad generalization of the uniform measure on permutations of size n, which corresponds to $\mu = \mathrm{Leb}_{[0,1]^2}$, the Lebesgue measure on $[0,1]^2$. Such models have been considered in the literature under various perspectives, see e.g. [DZ95, BDMW23, Dub23a, Dub23b, Sjö23].

In the current paper, we are interested in the binary search tree $\mathcal{T}\langle\sigma_\mu^n\rangle$ of this random permutation model. Since we will be interested only in the shape of this tree (height, subtree size convergence), we may and will equivalently consider the tree $\mathcal{T}\langle\mathcal{P}_\mu^n\rangle$ instead of $\mathcal{T}\langle\sigma_\mu^n\rangle$.

1.3. **First main result: universal behavior of the BST height.** For a (labeled binary) tree $\mathcal{T}$, we denote by $h(\mathcal{T})$ its height, i.e. the maximal distance from a leaf to the root. As mentioned in Section 1.1, Devroye [Dev86] proved that for uniform random permutations σ^n of size n, the quantity $h(\mathcal{T}\langle\sigma^n\rangle)/\log n$ converges in probability and in L^p (for all $p \geq 1$) to a constant c^*, defined as the unique solution to $c\log(2e/c) = 1$ with $c \geq 2$. Our first result gives a sufficient condition on a permuton μ, under which the same result holds for the height of $\mathcal{T}\langle\mathcal{P}_\mu^n\rangle$. In the following, a permuton μ is said to satisfy assumption (A1) if μ has a bounded density ρ on the whole unit square $[0,1]^2$, which is continuous and positive on a neighborhood of $\{0\} \times [0,1]$.

Theorem 1.1 (Universality of BST height for permuton samples)**.** *Let μ be a permuton satisfying assumption* (A1)*. Then, as n goes to infinity, the following convergence holds in probability and in L^p for any $p \geq 1$:*

$$\frac{h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right)}{c^* \log n} \longrightarrow 1\,.$$

The constant c^* in the above theorem is the same as in the uniform case, i.e. the unique solution to $c \log(2e/c) = 1$ with $c \geq 2$.

Let us comment on the assumption (A1). A natural sufficient condition is that the density ρ is continuous on the whole square $[0,1]^2$ and positive on the left edge $\{0\} \times [0,1]$. In Section 5.1, we will see that this positivity assumption cannot be skipped. Indeed we exhibit, for any $\delta > 0$, a permuton μ_δ with a continuous density vanishing on $\{0\} \times [\frac{1}{2}, 1]$ such that $\mathcal{T}\langle \mathcal{P}_\mu^n \rangle$ has height at least $\Theta(n^{(1-\delta)/2})$ with probability tending to 1.

We also provide an example of a permuton μ'_β with density 1 on the band $[0, \beta] \times [0,1]$ and for which, for any $\varepsilon > 0$, the tree $\mathcal{T}\langle \mathcal{P}_{\mu'_\beta}^n \rangle$ has height at least $\frac{1-\beta}{\beta+\varepsilon} \log n$ with high probability. Choosing β and ε such that $(1-\beta)/(\beta+\varepsilon) > c^*$ gives us an example of permuton which has a positive continuous density on a band $[0, \beta] \times [0,1]$, but for which the conclusion of Theorem 1.1 fails. Hence the existence of a density on the whole square in assumption (A1) is needed.

On the other hand, we could not construct a permuton μ such that $h\left(\mathcal{T}\langle \mathcal{P}_\mu^n \rangle\right)$ is asymptotically smaller than $c^* \log n$ with a non-vanishing probability. This leads us to the following conjecture.

Conjecture 1. *For any permuton μ and $\varepsilon > 0$, one has*

$$\lim_{n \to \infty} \mathbb{P}\left[\frac{h\left(\mathcal{T}\langle \mathcal{P}_\mu^n \rangle\right)}{\log n} < c^* - \varepsilon\right] = 0.$$

As evidence supporting Conjecture 1, we show that the statement holds for permutons with a density assumed to be continuous and positive around a single point $(0, y)$ on the left edge of the unit square; see Proposition 5.3.

Remark 1. For Mallows random permutations $\sigma_{n,q}$ with parameter q, it is known that there is a phase transition for $q = 1 + \beta n^{-1} + o(n^{-1})$, for which the model has a non-trivial permuton limit ν_β depending on β [Sta09]. This limiting permuton ν_β has a continuous density, which is nonzero on the whole square $[0,1]^2$. From Theorem 1.1, we know that $\mathcal{T}\langle \sigma_{\nu_\beta}^n \rangle$ has height $(c^* + o(1)) \log n$. Besides it was proved in [ABC21] that, in this regime, $\mathcal{T}\langle \sigma_{n,q} \rangle$ has height $(c^* + o(1)) \log n$ as well. None of these results imply the other one since $\sigma_{n,q}$ and $\sigma_{\nu_\beta}^n$ have different distributions, but these two regimes can be coupled in a rather strong way (see [MS13], where such a coupling is constructed to study the longest increasing subsequence), and it would have been surprising that the heights of their BSTs behave in different ways.

Remark 2. There is a natural way to construct the sequence of permutations $(\sigma_\mu^n)_{n \geq 1}$ on the same probability space: we consider a single infinite sequence $((x_i, y_i))_{i \geq 1}$ of i.i.d. points with law μ, and construct each σ_μ^n using the first n points of this sequence. We may wonder whether the convergence in Theorem 1.1 holds almost surely for this construction. We do not know whether this is the case, even for $\mu = \text{Leb}_{[0,1]^2}$. It has been shown by Pittel in [Pit84] that, if $Y_1, Y_2, \ldots$ is an infinite sequence of uniform random variables in $[0,1]$, then $h(\mathcal{T}\langle Y_1, \ldots, Y_n \rangle)/\log n$ converges a.s. to a constant β. Combining with the above-mentioned result of Devroye, the constant β must be equal to c^*, as noted in [Dev86, Section 5]. However, the sequence $(h(\mathcal{T}\langle Y_1, \ldots, Y_n \rangle))_{n \geq 1}$ considered by Pittel does not have the same distribution as $\left(h\left(\mathcal{T}\langle \mathcal{P}_\mu^n \rangle\right)\right)_{n \geq 1}$ for $\mu = \text{Leb}_{[0,1]^2}$. Indeed, while both have the same distribution for each n, the couplings between different values of n are different: in Pittel's model, the new element Y_{n+1} is always added at the end of the sequence $Y_1, \ldots, Y_n$, while in our model, the new element of $Y_{(1)}, \ldots, Y_{(n+1)}$ is added in a uniformly random position in $Y_{(1)}, \ldots, Y_{(n)}$ ($Y_{(i)}$ depends implicitly on n, not only on i).

1.4. **Second main result: subtree size convergence of the BSTs.** In this section, we state a limit theorem for $\mathcal{T}\langle \mathcal{P}_\mu^n \rangle$ (under a mild assumption on μ) in the sense of the subtree size convergence recently introduced by Grübel [Grü23]. We first recall this notion of convergence.

From now on, we identify nodes in a binary tree with finite words in the alphabet $\{0, 1\}$ as follows: the empty word $\varnothing$ corresponds to the root, and for a node v encoded by w, the words $w0$ and $w1$

obtained by appending 0 or 1 to w encode respectively the left and right children of v. Moreover, we let $\mathbb{V} = \{0,1\}^*$ be the set of all finite words on $\{0,1\}$, representing all nodes of the complete infinite binary tree. With this notation, a labeled tree is identified with a function from a subset of $\mathbb{V}$ to $\mathbb{R}$, where the domain of the function is the set of nodes in the tree, and a node is mapped to its label. In particular, $\mathcal{T}(v)$ denotes the label of the node v in $\mathcal{T}$. We also write $v \in \mathcal{T}$ to indicate that the node v is in the tree $\mathcal{T}$.

Given a finite (potentially labeled) tree $\mathcal{T}$ and a node $v \in \mathbb{V}$, we let

$$t(\mathcal{T}, v) := \frac{1}{|\mathcal{T}|} \left| \left\{ u \in \mathcal{T} : v \preceq u \right\} \right|,$$

where $v \preceq u$ means that v is a prefix of u. In words, $t(\mathcal{T}, v)$ is the proportion of nodes in $\mathcal{T}$ which are descendants of v.

Further write Ψ for the set of functions $\psi : \mathbb{V} \to [0,1]$ such that $\psi(\varnothing) = 1$ and such that, for any $v \in \mathbb{V}$, we have $\psi(v) = \psi(v0) + \psi(v1)$. Then a sequence of binary trees $(\mathcal{T}^n)_{n \in \mathbb{N}}$ is said to converge to a function $\psi \in \Psi$ if and only if, for any $v \in \mathbb{V}$, the quantity $t(\mathcal{T}^n, v)$ converges to $\psi(v)$. If that is the case, we write $\mathcal{T}_n \xrightarrow{\text{ssc}} \psi$, and refer to this as *subtree size convergence*.

We now define two important objects for the subtree size convergence of BSTs of permuton samples. For any complete infinite BST $\mathcal{T} : \mathbb{V} \to (0,1)$ with labels in $(0,1)$, we define $\mathcal{T}_{\text{left}} : \mathbb{V} \to \mathbb{R}$ as follows. First of all, if v consists only of zeros, we let $\mathcal{T}_{\text{left}}(v) = 0$. Now, given that $v = v'10^k$ for some $k \geq 0$, let $\mathcal{T}_{\text{left}}(v) = \mathcal{T}(v')$. Informally, $\mathcal{T}_{\text{left}}(v)$ is the right-most ancestor of v to its left. Define similarly $\mathcal{T}_{\text{right}}$ such that $\mathcal{T}_{\text{right}}(v) = 1$ if v consists only of ones, and $\mathcal{T}_{\text{right}}(v) = \mathcal{T}(v')$ whenever $v = v'01^k$ for some $k \geq 0$. In words, $\mathcal{T}_{\text{right}}(v)$ is the left-most ancestor of v to its right. See Figure 3 for an illustration of $\mathcal{T}_{\text{left}}$ and $\mathcal{T}_{\text{right}}$. We note that these definitions imply that $\mathcal{T}_{\text{left}}(v) < \mathcal{T}(v) < \mathcal{T}_{\text{right}}(v)$ for any $v \in \mathbb{V}$.
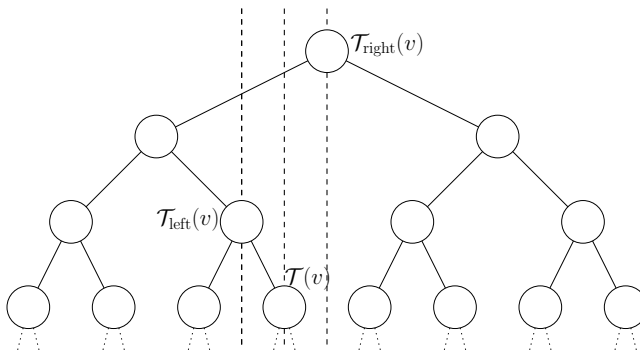

FIGURE 3. Representation of $\mathcal{T}_{\text{right}}$ and $\mathcal{T}_{\text{left}}$ on a labeled BST, for the node $v = 011$.

Given a probability measure m on $[0,1]$ without atoms, write $\psi_m \in \Psi$ for the following random object. First, let $Y = (Y_1, Y_2, \ldots)$ be an infinite sequence of independent random variables distributed according to m and write $\mathcal{T}^m = \mathcal{T}\langle Y \rangle$ for the corresponding (infinite) BST. Then, let $\psi_m = \mathcal{T}^m_{\text{right}} - \mathcal{T}^m_{\text{left}}$. This is well-defined, since $\mathcal{T}^m$ is complete with probability 1 (this follows from [Dev86, Theorem 6.1] and the fact that the shape of $\mathcal{T}^m = \mathcal{T}\langle Y \rangle$ is independent of m). Further note that this object indeed belongs a.s. to Ψ, since $\mathcal{T}_{\text{left}}(v0) = \mathcal{T}_{\text{left}}(v)$, $\mathcal{T}_{\text{left}}(v1) = \mathcal{T}(v)$, $\mathcal{T}_{\text{right}}(v0) = \mathcal{T}(v)$, and $\mathcal{T}_{\text{right}}(v1) = \mathcal{T}_{\text{right}}(v)$.

To illustrate this construction, take $m = 2x dx$, where dx is the Lebesgue measure on $[0,1]$. Here is an i.i.d. sample of size 10 from m: $(0.73, 0.33, 0.75, 0.35, 0.68, 0.28, 0.72, 0.87, 0.25, 0.67)$. Its associated BST, which is the top part of the BST $\mathcal{T}^m = \mathcal{T}\langle Y \rangle$ associated with an infinite sample, is given in Figure 4, left. The associated realization of the function ψ_m is then given in Figure 4, right. For instance, for the node $v = 011$ as chosen in Figure 3, we have $\psi_m(v) = \mathcal{T}^m_{\text{right}}(v) - \mathcal{T}^m_{\text{left}}(v) = \mathcal{T}^m(\varnothing) - \mathcal{T}^m(01) = 0.73 - 0.35 = 0.38$, as written in Figure 4.
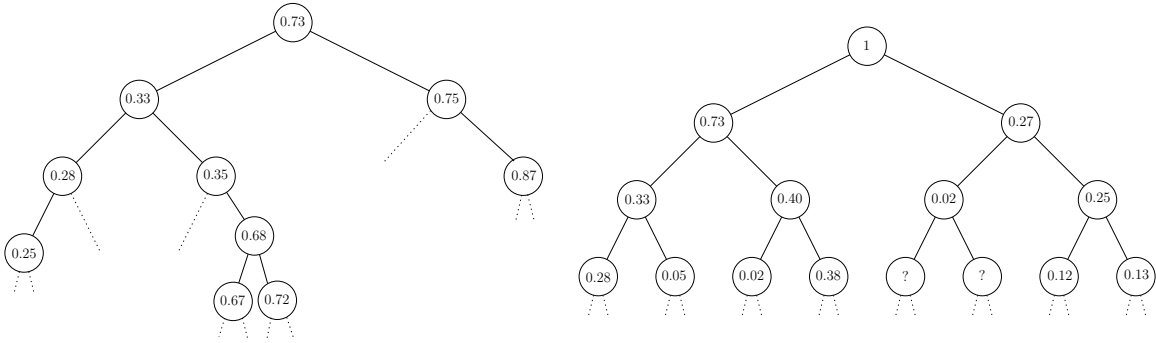
FIGURE 4. Example of realizations of $\mathcal{T}^m$ and ψ_m with $m = 2xdx$. Note that we do not have enough data to compute two of the values of ψ_m on nodes in the third level; those nodes are marked with question marks.

We can now state our second main result. A permuton is said to satisfy assumption (A2) if there exists a probability measure μ_0 on $[0, 1]$ without atoms such that

$$(2) \qquad \frac{1}{x}\mu\big([0, x] \times \cdot\big) \xrightarrow[x \to 0]{} \mu_0,$$

where the convergence occurs according to the weak topology of probability measures on $[0, 1]$. Assumption (A2) is weaker than (A1): in particular, (A2) holds whenever μ admits a continuous density on a neighborhood of $\{0\} \times [0, 1]$, without any further assumption on that density.

Theorem 1.2 (Subtree-size convergence of BSTs of permuton samples). *Let μ be a permuton satisfying* (A2) *and let μ_0 be defined by* (2). *Then, we have the following convergence in distribution for the subtree size topology:*

$$\mathcal{T}\langle \mathcal{P}_\mu^n \rangle \xrightarrow{\text{ssc}} \psi_{\mu_0}.$$

It is interesting to see that the limit depends on μ only through μ_0. The assumption that μ_0 does not have atoms is important. A first difficulty when μ_0 has some atom is that the BST $\mathcal{T}\langle Y_1, Y_2, \dots \rangle$ where $Y_1, Y_2, \dots$, are i.i.d. random variables with distribution μ_0 is ill-defined since some of the Y_i's are equal. We can also see that, in this case, the limit of $\mathcal{T}\langle \mathcal{P}_\mu^n \rangle$ may not depend only on μ_0. Indeed, consider the permuton μ^1 (resp. μ^2) supported by the set $y \equiv \frac{1}{2} + x$ modulo 1 (resp. $y \equiv \frac{1}{2} - x$ modulo 1). They both satisfy (2) with $\mu_0 = \delta_{1/2}$. But it is easy to see that the trees $\mathcal{T}\langle \mathcal{P}_{\mu^1}^n \rangle$ and $\mathcal{T}\langle \mathcal{P}_{\mu^2}^n \rangle$ have different limits in the sense of subtree size convergence: in particular, one has

$$\lim_{n \to \infty} t(\mathcal{T}\langle \mathcal{P}_{\mu^1}^n \rangle, 11) = \tfrac{1}{2}, \quad \text{while} \quad \lim_{n \to \infty} t(\mathcal{T}\langle \mathcal{P}_{\mu^2}^n \rangle, 11) = 0.$$

1.5. **Decomposition of BSTs and proof strategies.** In this section we present a useful decomposition of the BSTs drawn from permuton samples, and provide an overview of the proofs of Theorems 1.1 and 1.2.

Decomposing a BST from a permuton sample. A basic idea in this work consists in decomposing the BST drawn from a permuton sample, as a top tree to which hanging trees are attached. To this end, we first consider the K left-most points in the sample $\mathcal{P}_\mu^n$ (i.e. the K points with smallest x-coordinates). These K points are the first ones to be inserted in the construction of $\mathcal{T}\langle \mathcal{P}_\mu^n \rangle$ and are therefore inserted at the top of the tree. We will refer throughout the paper to the part of $\mathcal{T}\langle \mathcal{P}_\mu^n \rangle$ corresponding to these first K points as the *top tree*. The labels in the top tree correspond to the y-coordinates $y_{(1)}, \dots, y_{(K)}$ of these first K points. Now, consider the subdivision $I_1, \dots, I_{K+1}$ of $[0, 1]$ induced by these numbers $y_{(1)}, \dots, y_{(K)}$. In the construction of $\mathcal{T}\langle \mathcal{P}_\mu^n \rangle$, further points (x, y) will be inserted between some pair of consecutive vertices in the top tree, depending on the index j such that $y \in I_j$. Hence the tree $\mathcal{T}\langle \mathcal{P}_\mu^n \rangle$

is obtained by grafting to the top tree one subtree for each interval I_j; see Figure 5. These trees will be refered to as the *hanging trees*.
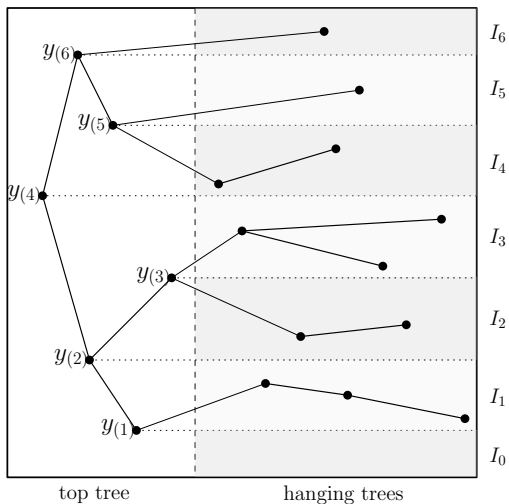

FIGURE 5. A sample of points and its associated BST, decomposed as top and hanging trees (for $K = 6$). The BST has been rotated of 90 degrees to the left, so that it can be drawn directly on the set of points.

Subtree size convergence. From this decomposition, the proof of the subtree size convergence is relatively simple. We take K large, but independent of n. First of all we prove that, under the regularity assumption (A2), the first K points look like i.i.d. random variables sampled from the left distribution μ_0 (see Proposition 2.1 and Corollary 2.2). Moreover, since a permuton μ has by definition uniform projections, the proportion of points in each horizontal band seen in Figure 5 is asymptotically given by the height of the band. By choosing K large enough, this gives us the proportion of nodes in all subtrees until any given depth, thus proving the subtree size convergence (see Lemma 2.3 and the proof of Theorem 1.2).

Height universality. To prove Theorem 1.1, we use the same decomposition of the BST into a top tree and hanging trees, but this time we take $K = \beta n + o_{\mathbb{P}}(n)$ with β small but independent from n. The heights of the top tree and hanging trees are controlled via different approaches.

For the top tree, our assumption (A1) (specifically the continuity and positivity of the density near the left edge) ensures the following: for β small enough, the density restricted to the left vertical band $[0, \beta] \times [0, 1]$ is close, up to a multiplicative factor, to a density depending only on the y-coordinate. It is easy to see that if the density only depended on y, then the associated permutation would be uniformly distributed, and thus the BST would have height $(c^* + o_{\mathbb{P}}(1)) \log K = (c^* + o_{\mathbb{P}}(1)) \log n$. We then use comparison arguments to prove that the top tree also has height $(c^* + o_{\mathbb{P}}(1)) \log n$ under assumption (A1). Such comparison arguments need to be handled carefully, since adding a single point to a point set can halve the height of the associated BST (see Remark 3). On the other hand, we show that removing a point cannot decrease this height by more than 1 (Lemma 3.2), and use this to control the height of the top tree.

It remains to argue that the hanging trees all have height $o_{\mathbb{P}}(\log n)$. For $K = \beta n + o_{\mathbb{P}}(n)$, the horizontal bands in Figure 5 contain $\mathcal{O}(1)$ points on average, but the largest number of points in a band is actually $\mathcal{O}(\log n)$ (see Proposition 4.4). The hanging trees are themselves BSTs of random point sets, and therefore we expect them to have height $\mathcal{O}(\log \log n) \ll \log n$. However, there are $\mathcal{O}(n)$ hanging trees, and we need all of them to have height $o(\log n)$. To prove this, we need good deviation estimates on the fact that the BST of a point set has a height negligible compared to its size.

Such estimates are provided by Devroye for uniform BSTs [Dev86, Lemma 3.1], but the monotonicity properties of BSTs are not good enough to use direct comparison arguments. We solve this difficulty by comparing, for any point set, the height of its BST and the length of its longest monotone subsequences, and then by using monotonicity properties and deviation bounds for the latter (see Lemma 3.5 and Corollary 3.7).

Fixed number of points and Poisson point processes. As in many results considering point sets, it is often more convenient to work with a Poisson number $N \sim \text{POISSON}(n)$ of points, instead of a fixed number n. Indeed, the point set then becomes a Poisson point process on $[0,1]^2$ with intensity $n\mu$ and gains useful independence properties: conditionally given the labels $y_{(1)}, \ldots, y_{(k)}$ in the top tree, the hanging trees are independent from each other.

In Theorem 3.4 we provide a general de-Poissonization result, relating the asymptotic height of a BST constructed from a Poisson point process with intensity $n\mu$ to that of n i.i.d. points with law μ. The proof of this result uses the comparison lemma already mentioned above (Lemma 3.2), and is made difficult by the fact that we only have a control in one direction (recall that adding a single point to a point set can halve the height of the associated BST).

1.6. **Basic probabilistic facts and notation.** Throughout the paper, "with high probability" (or w.h.p. for short) means "with probability tending to 1, as n tends to ∞". We also use the notation $X_n = o_{\mathbb{P}}(Y_n)$ to say that X_n/Y_n converges to 0 in probability.

We start by stating a useful lemma, which compares the size of random subsets to binomial random variables. We write $X \preceq Y$ (resp. $X \succeq Y$) to denote that X is stochastically smaller (resp. larger) than Y.

Lemma 1.3. *Let m be an integer and let $i, j < m$. Let I, J be subsets of $\{1, \ldots, m\}$ of respective sizes i, j, where I is fixed and J is uniformly random. Then the following stochastic domination holds:*

$$|I \cap J| \preceq \text{BINOMIAL}\left(i, \frac{j}{m-i}\right)$$

where by convention the law $\text{BINOMIAL}(n, p)$ with $p > 1$ is the Dirac law at n.

Proof. By symmetry, $|I \cap J|$ would have the same distribution if J were fixed and I taken uniformly at random among subsets of size i of $\{1, \ldots, m\}$. In this situation, I can be constructed by uniformly picking i elements of $\{1, \ldots, m\}$ without replacement. Each picked element has probability at most $\frac{j}{m-i}$ of being in J, which proves the lemma. □

Next, we record some well-known asymptotic estimates for Poisson random variables with large parameters. If N is a $\text{POISSON}(n)$ distributed random variable, then N/n converges to 1 in distribution and in moments: i.e., for any fixed integer $p \geq 1$, we have

$$(3) \qquad\qquad\qquad \mathbb{E}[N^p] = n^p(1 + o(1)).$$

Furthermore, we have the following tail estimates, easily obtained via Chernoff bounds.

Lemma 1.4. *Let $\lambda > 0$ and X be a $\text{POISSON}(\lambda)$ distributed random variable. Then, for $x \geq \lambda$, we have*

$$\mathbb{P}[X \geq x] \leq \left(\frac{e\lambda}{x}\right)^x e^{-\lambda},$$

while, for $x \leq \lambda$, it holds that

$$\mathbb{P}[X \leq x] \leq \left(\frac{e\lambda}{x}\right)^x e^{-\lambda}.$$

Proof. Assume $x \geq \lambda$. For $\theta > 0$, we have

$$\mathbb{P}[X \geq x] \leq \frac{\mathbb{E}[e^{\theta X}]}{e^{\theta x}} = e^{\lambda e^{\theta} - \lambda - \theta x},$$

and the first inequality in the lemma follows by setting $e^{\theta} = x/\lambda$. The second one is proved similarly. □

We conclude this section with a variant of the classical Glivenko–Cantelli theorem for triangular arrays.

Proposition 1.5. *Let μ be a probability measure with distribution function $F(x) := \mu((-\infty, x])$, and a finite fourth moment. For each $n \geq 1$, let $(X_{i,n})_{1 \leq i \leq n}$ be i.i.d. random variables with common distribution μ and let*

$$F_n(x) := \frac{1}{n} \left| \{ i \leq n : X_{i,n} \leq x \} \right|$$

be their empirical distribution function. Then, a.s. it holds that F_n converges uniformly to F.

Proof. A classical fourth moment computation, together with Borel–Cantelli lemma – see, *e.g.*, [Bil12, Theorem 6.1] – shows that, for any fixed x, $F_n(x)$ converges a.s. to $F(x)$. The rest of the proof is similar to that of the classical Glivenko–Cantelli theorem which considers a single sequence X_i of i.i.d. random variables instead of a triangular array, but does not require a fourth moment condition; see, *e.g.*, [Bil12, Theorem 20.6]. $\square$

2. SUBTREE SIZE CONVERGENCE

2.1. **Convergence of the first elements.** In the following proposition, μ is a permuton and for each n, we let N be a POISSON (n) distributed random variable. We drop the subscript μ for clarity and let $\mathcal{P}^N = \left\{ \left(X_i^N, Y_i^N \right) \right\}_{i \leq N}$ be i.i.d. random variables with distribution μ; equivalently $\mathcal{P}^N$ is a Poisson point process on $[0,1]^2$ with intensity $n\mu$. Also let $\left(\left(X_{(i)}^N, Y_{(i)}^N \right) \right)_{i \leq N}$ be its reordering such that $X_{(1)}^N < \cdots < X_{(N)}^N$.

Proposition 2.1. *Assume that there exists a measure μ_0 satisfying (2), that is*

$$\lim_{x \to 0} \frac{1}{x} \mu\big([0, x] \times \cdot \big) = \mu_0 \,,$$

where the convergence occurs according to the weak topology. Then, for any fixed $K \geq 1$, we have the following convergence in distribution:

$$\left(Y_{(1)}^N, \ldots, Y_{(K)}^N \right) \xrightarrow[n \to \infty]{} (Y_k)_{1 \leq k \leq K} \,,$$

where $(Y_k)_{1 \leq k \leq K}$ is a sequence of K i.i.d. random variables distributed according to μ_0.

Note that, at this stage, we do not need to assume that μ_0 has no atoms as in (A2).

Proof. Let $I_1, \ldots, I_K$ be intervals of $[0,1]$ whose boundaries contain no atom of μ_0. We aim to show that:

$$(4) \qquad \mathbb{P}\left[\forall k \leq K, \ Y_{(k)}^N \in I_k \right] \xrightarrow[n \to \infty]{} \mu_0(I_1) \ldots \mu_0(I_K).$$

Fix $\varepsilon > 0$ and set $L := \lfloor 1/\varepsilon^3 \rfloor$. By taking ε small enough, we can assume that $L \geq K$. Using the assumption (2) and the Portmanteau theorem, there exists $x_0 > 0$ such that for any $x \leq x_0$ and any $k \leq K$:

$$(5) \qquad \left| \frac{1}{x} \mu\big([0, x) \times I_k \big) - \mu_0 \left(I_k \right) \right| \leq \varepsilon / L \,,$$

Set $n_0 := \lceil \frac{1}{\varepsilon x_0} \rceil$ and consider any $n \geq n_0$. For all $0 \leq i \leq L - 1$ and $1 \leq k \leq K$, define the following blocks and columns:

$$B_{i,k} := \left[\frac{i}{L\varepsilon n}, \frac{i+1}{L\varepsilon n} \right) \times I_k \quad \text{and} \quad C_i := \left[\frac{i}{L\varepsilon n}, \frac{i+1}{L\varepsilon n} \right) \times [0, 1].$$

Using $\frac{1}{\varepsilon n} \leq x_0$ and (5), we have for each $0 \leq i \leq L - 1$ and $1 \leq k \leq K$:

$$
(6) \qquad \left| \mu\left(B_{i,k}\right) - \frac{1}{L\varepsilon n} \mu_0(I_k) \right|
$$

$$
\leq \left| \mu\left(\left[0, \frac{i+1}{L\varepsilon n}\right) \times I_k \right) - \frac{i+1}{L\varepsilon n} \mu_0(I_k) \right| + \left| \frac{i}{L\varepsilon n} \mu_0(I_k) - \mu\left(\left[0, \frac{i}{L\varepsilon n}\right) \times I_k \right) \right|
$$

$$
\leq \frac{i+1}{L\varepsilon n} \frac{\varepsilon}{L} + \frac{i}{L\varepsilon n} \frac{\varepsilon}{L}
$$

$$
\leq \frac{2}{nL}.
$$

Now define the events

$$
E_1 := \left\{ \forall i \leq L - 1, \ \left|\mathcal{P}^N \cap C_i\right| \leq 1 \right\} \quad \text{and} \quad E_2 := \left\{ \left| \bigcup_{0 \leq i \leq L-1} C_i \cap \mathcal{P}^N \right| \geq K \right\}.
$$

In words, E_1 is the event that each one of the first L columns of width $1/L\varepsilon n$ contains at most one point, while E_2 is the event that there are at least K points in total in all those columns, that is in the set $[0, 1/\varepsilon n) \times [0, 1]$. Each $\left|\mathcal{P}^N \cap C_i\right|$ follows a Poisson law of parameter $n\mu(C_i) = \frac{1}{L\varepsilon}$, thus

$$
1 - \mathbb{P}[E_1] = \mathbb{P}\left[\exists i \leq L - 1, \ |\mathcal{P}^N \cap C_i| \geq 2 \right] \leq L\left(1 - e^{-\frac{1}{L\varepsilon}} - \frac{1}{L\varepsilon} e^{-\frac{1}{L\varepsilon}} \right).
$$

Recalling that $L = \lfloor 1/\varepsilon^3 \rfloor$, it follows that

$$
\mathbb{P}[E_1] \geq 1 - \mathcal{O}\left(\frac{1}{L\varepsilon^2} \right) = 1 - \mathcal{O}(\varepsilon),
$$

where the constant in $\mathcal{O}(\cdot)$ is independent from n and ε. Likewise $\left|(\cup_i C_i) \cap \mathcal{P}^N\right|$ follows a Poisson law of parameter $1/\varepsilon$, so $\mathbb{P}[E_2] \geq 1 - o(1)$ as $\varepsilon \to 0$, for any fixed K and uniformly in n. We conclude that the event $E := E_1 \cap E_2$ satisfies $\mathbb{P}[E] \geq 1 - \delta(\varepsilon)$, where $\lim_{\varepsilon \to 0} \delta(\varepsilon) = 0$ and δ does not depend on n.

Under the event E, each column C_i contains at most one point, but all the columns together contain at least K points. This means that the column indices $i_1, \ldots, i_K$ of the K left-most points of $\mathcal{P}^N$ satisfy $0 \leq i_1 < \ldots < i_K \leq L - 1$. Thus, we get

$$
\mathbb{P}\left[\left\{ \forall k \leq K, \ Y_{(k)}^N \in I_k \right\} \cap E \right] = \sum_{0 \leq i_1 < \cdots < i_K \leq L-1} \mathbb{P}\left[\left\{ \forall k \leq K, \ \left(X_{(k)}^N, Y_{(k)}^N \right) \in B_{i_k, k} \right\} \cap E \right].
$$

The latter event is equivalent to the conjunction of four facts: i) for each $k \leq K$, $B_{i_k, k}$ contains exactly 1 point from $\mathcal{P}^N$; ii) for each $k \leq K$, the remainder $C_{i_k} \setminus B_{i_k, k}$ of the column C_{i_k} contains no point; iii) the columns C_i with $i \leq i_K$ and $i \notin \{i_1, \ldots, i_K\}$ are empty; and, iv) each column C_i with $i_K < i \leq L - 1$ contains at most 1 point. In a Poisson point process, the number of points in disjoint sets are independent Poisson random variables, so we get

$$
\mathbb{P}\left[\left\{ \forall k \leq K, \ Y_{(k)}^N \in I_k \right\} \cap E \right] = \sum_{i_1 < \cdots < i_K} \left[\prod_{k=1}^{K} n\mu\left(B_{i_k, k}\right) e^{-n\mu(B_{i_k,k})} e^{-n(\mu(C_{i_k}) - \mu(B_{i_k,k}))} \right.
$$

$$
\left. \cdot \prod_{\substack{i \leq i_K \\ i \notin \{i_1, \ldots, i_K\}}} e^{-n\mu(C_i)} \prod_{i=i_K+1}^{L-1} e^{-n\mu(C_i)} \left(1 + n\mu(C_i)\right) \right].
$$

Combining all the exponential terms simplifies to $e^{-n\mu([0, \frac{1}{n\varepsilon}) \times [0,1])} = e^{-1/\varepsilon}$. Moreover for each $i > i_K$, we have $\mu(C_i) = \frac{1}{L\varepsilon n}$. Therefore, we get

$$
\mathbb{P}\left[\left\{ \forall k \leq K, \ Y_{(k)}^N \in I_k \right\} \cap E \right] = n^K e^{-1/\varepsilon} \sum_{i_1 < \cdots < i_K} \left(1 + \frac{1}{L\varepsilon} \right)^{L - i_K - 1} \prod_{k=1}^{K} \mu\left(B_{i_k, k}\right).
$$

Using (6), we deduce that

(7)
$$\mathbb{P}\left[\left\{\forall k \leq K,\, Y_{(k)}^N \in I_k\right\} \cap E\right]$$

$$\leq n^K e^{-1/\varepsilon} \sum_{i_1 < \cdots < i_K} \left(1 + \frac{1}{L\varepsilon}\right)^{L-i_K-1} \prod_{k=1}^K \left(\frac{1}{L\varepsilon n}\mu_0(I_k) + \frac{2}{nL}\right)$$

$$\leq c(\varepsilon) \prod_{k=1}^K \left(\mu_0(I_k) + 2\varepsilon\right)$$

where

(8)
$$c(\varepsilon) := \left(\frac{1}{L\varepsilon}\right)^K e^{-1/\varepsilon} \sum_{i_K=K-1}^{L-1} \left(1 + \frac{1}{L\varepsilon}\right)^{L-i_K-1} \binom{i_K}{K-1}.$$

Likewise

(9)
$$\mathbb{P}\left[\left\{\forall k \leq K,\, Y_{(k)}^N \in I_k\right\} \cap E\right] \geq c(\varepsilon) \prod_{k=1}^K \left(\left(\mu_0(I_k) - 2\varepsilon\right)_+\right),$$

where we use the notation $x_+ = \max(x, 0)$. Moreover this reasoning can be applied with each I_k being the whole interval $[0, 1]$, yielding:

$$c(\varepsilon)\,(1 - 2\varepsilon)^K \leq \mathbb{P}(E) \leq c(\varepsilon)\,(1 + 2\varepsilon)^K$$

Since $1 - \delta(\varepsilon) \leq \mathbb{P}(E) \leq 1$, we get the following bounds for $c(\varepsilon)$:

$$\frac{1 - \delta(\varepsilon)}{(1 + 2\varepsilon)^K} \leq c(\varepsilon) \leq \frac{1}{(1 - 2\varepsilon)^K}.$$

In particular, this implies that $\lim_{\varepsilon \to 0} c(\varepsilon) = 1$. Using (7), we now deduce that, for any $n \geq n_0$:

$$\mathbb{P}\left[\forall k \leq K,\, Y_{(k)}^N \in I_k\right] \leq \mathbb{P}\left[\left\{\forall k \leq K,\, Y_{(k)}^N \in I_k\right\} \cap E\right] + \delta(\varepsilon) \leq c(\varepsilon) \prod_{k=1}^K \left(\mu_0(I_k) + 2\varepsilon\right) + \delta(\varepsilon),$$

and likewise with (9):

$$\mathbb{P}\left[\forall k \leq K,\, Y_{(k)}^N \in I_k\right] \geq \mathbb{P}\left[\left\{\forall k \leq K,\, Y_{(k)}^N \in I_k\right\} \cap E\right] \geq c(\varepsilon) \prod_{k=1}^K \left(\left(\mu_0(I_k) - 2\varepsilon\right)_+\right).$$

Both bounds converge to $\mu_0(I_1) \cdots \mu_0(I_K)$ when $\varepsilon \to 0$, proving (4). This concludes the proof of the proposition. $\qquad\square$

We now "de-Poissonize" the previous result. We use the notation of Section 1.2, namely $(X_i^n, Y_i^n)_{i \leq n}$ is an i.i.d. sample of fixed size n and common distribution μ, and $(X_{(i)}^n, Y_{(i)}^n)_{i \leq n}$ is its reordering with increasing x-coordinates.

Corollary 2.2. *Assume that there exists a measure μ_0 such that*

$$\lim_{x \to 0} \frac{1}{x}\mu\big([0, x] \times \cdot\big) = \mu_0,$$

where the convergence occurs according to the weak topology. Then, for any fixed $K \geq 1$, we have the following convergence in distribution:

$$\left(Y_{(1)}^n, \ldots, Y_{(K)}^n\right) \xrightarrow[n \to \infty]{} (Y_k)_{1 \leq k \leq K},$$

where $(Y_k)_{1 \leq k \leq K}$ is a sequence of K i.i.d. random variables distributed according to μ_0.

Proof. For each $n \in \mathbb{N}$, let $\mathcal{P}^N = \left((X_1^N, Y_1^N), \ldots, (X_N^N, Y_N^N) \right)$ be a Poisson point process of intensity $(n + n^{2/3})\mu$, listed in a uniform random order. Since the number N of points follows a POISSON $(n + n^{2/3})$ law, for large n, it has fluctuations of order $\sqrt{n}$ around its mean value $n + n^{2/3}$. In particular, the event $\{N \geq n\}$ happens w.h.p. as n goes to infinity. Conditionally under this event, $\{(X_1^N, Y_1^N), \ldots, (X_n^N, Y_n^N)\}$ is a family of n i.i.d. random points distributed under μ. We denote by $(X_{(1)}^N, Y_{(1)}^N), \ldots, (X_{(N)}^N, Y_{(N)}^N)$ and $(X_{(1)}^n, Y_{(1)}^n), \ldots, (X_{(n)}^n, Y_{(n)}^n)$ the reorderings, by increasing x-coordinate, of $\{(X_1^N, Y_1^N), \ldots, (X_N^N, Y_N^N)\}$ and $\{(X_1^N, Y_1^N), \ldots, (X_n^N, Y_n^N)\}$. Let τ be the unique permutation of size N satisfying $X_{(i)}^N = X_{\tau(i)}^N$ for all $1 \leq i \leq N$. Since τ is uniformly random and since $N \leq n + 2n^{2/3}$ w.h.p., the event $\{\tau(1) \leq n, \ldots, \tau(K) \leq n\}$ happens w.h.p. as n goes to infinity. Informally, this event means that the K left-most points of the whole Poisson point process $\mathcal{P}^N$ belong to the subset of its first n points. Conditionally under this event, one has:

$$\left((X_{(1)}^n, Y_{(1)}^n), \ldots, (X_{(K)}^n, Y_{(K)}^n) \right) = \left((X_{(1)}^N, Y_{(1)}^N), \ldots, (X_{(K)}^N, Y_{(K)}^N) \right).$$

Corollary 2.2 then follows from Proposition 2.1. $\square$

2.2. Proof of subtree size convergence.

Recall from Section 1.4 that for a finite tree $\mathcal{T}$ and a node u in $\mathbb{V}$, we denote by $t(\mathcal{T}, u)$ the proportion of nodes of $\mathcal{T}$ which are descendants of u (including u itself). In a binary search tree, this can be computed as follows.

Lemma 2.3. *Let $y_1, \ldots, y_n$ be distinct numbers and let $\mathcal{T} := \mathcal{T}\langle y_1, \ldots, y_n \rangle$ be the corresponding BST. Let u be a node in $\mathcal{T}$ and let k be such that $\mathcal{T}(u) = y_k$. Then we have*

$$(10) \qquad t(\mathcal{T}, u) = \frac{1}{|\mathcal{T}|} \left| \{y_k, \ldots, y_n\} \cap \left(\mathcal{T}_{\text{left}}(u), \mathcal{T}_{\text{right}}(u) \right) \right|,$$

where $\mathcal{T}_{\text{left}}$ and $\mathcal{T}_{\text{right}}$ are extended to finite binary search trees using the definition from Section 1.4.

Proof. Consider the iterative construction of $\mathcal{T}\langle y_1, \ldots, y_n \rangle$. A number y_i in the list is inserted in a node which is a strict descendant of u if

- the node u has been filled before, i.e. if $i > k$;
- the number y_i compares in the same way as y_k to all numbers $\mathcal{T}(u')$, where u' is an ascendant of u. This condition is equivalent to $y_i \in \left(\mathcal{T}_{\text{left}}(u), \mathcal{T}_{\text{right}}(u) \right)$, see Figure 3.

Hence, the numerator in (10) is indeed the number of descendants of u in T (including u, which corresponds to the label y_k). This proves the lemma. $\square$

Proof of Theorem 1.2. Let $\mathcal{T}^n := \mathcal{T}\langle \mathcal{P}_\mu^n \rangle$, skipping the dependency on μ in the notation. Since the subtree size topology is by definition the pointwise convergence of the function $(t(., u))_{u \in \mathbb{V}}$, we need to prove the convergence of finite-dimensional distributions. Namely, we need to prove that, for any $d \geq 1$ and $u_1, \ldots, u_d$ in $\mathbb{V}$, we have the following convergence in distribution as n tends to ∞:

$$(11) \qquad \left(t(\mathcal{T}^n, u_i) \right)_{i \leq d} \longrightarrow \left(\psi_\mu(u_i) \right)_{i \leq d}.$$

As usual, for each n, let $(X_1^n, Y_1^n), \ldots, (X_n^n, Y_n^n)$ be the i.i.d. random points in $[0,1]^2$ with common distribution μ used to construct $\mathcal{T}\langle \mathcal{P}_\mu^n \rangle$ and reorder them as $(X_{(1)}^n, Y_{(1)}^n), \ldots, (X_{(n)}^n, Y_{(n)}^n)$ such that $X_{(1)}^n < \cdots < X_{(n)}^n$.

From Corollary 2.2, we have the following convergence in distribution for the topology of pointwise convergence:

$$(12) \qquad \left(Y_{(k)}^n \right)_{k \geq 1} \xrightarrow[n \to \infty]{} (Y_k)_{k \geq 1},$$

where $Y_1, Y_2, \ldots$ is an infinite sequence of i.i.d. random variables with distribution μ_0. Using Skorohod's representation theorem [Bil99, Section 6], we might assume that the above convergence holds almost surely.

Since μ_0 has no atoms, the numbers $(Y_k)_{k \geq 1}$ are a.s. distinct. Moreover the tree $\mathcal{T}\langle Y_1, Y_2, \ldots \rangle$ has a.s. shape $\mathbb{V}$ (i.e. there is no empty branch). Consequently, a.s., there exists a (random) threshold K such that all nodes u_i belong to $\mathcal{T}\langle Y_1, \ldots, Y_K \rangle$. Using the convergence (12), we know that there exists

a (random) threshold n_0 such that for $n \geq n_0$, the relative order of $(Y^n_{(1)}, \ldots, Y^n_{(K)})$ is the same as that of $(Y_1, \ldots, Y_K)$. This implies that the trees $\mathcal{T}^n_K := \mathcal{T}\langle Y^n_{(1)}, \ldots, Y^n_{(K)}\rangle$ and $\mathcal{T}^\infty_K := \mathcal{T}\langle Y_1, \ldots, Y_K\rangle$ have the same shape T_K. Moreover, for any v in T_K, the values $\mathcal{T}^n_K(v)$ and $\mathcal{T}^\infty_K(v)$ correspond to $Y^n_{(i)}$ and Y_i respectively, for the *same* index i. Therefore, using again (12), we know that $\mathcal{T}^n_K(v)$ converges to $\mathcal{T}^\infty_K(v)$ (a.s. in the probability space created by the application of Skorohod's representation theorem).

Now, using Lemma 2.3 and the fact that each u_i is filled in $\mathcal{T}_n$ before step $K = \mathcal{O}_\mathbb{P}(1)$, we have that

$$t(\mathcal{T}^n, u_i) = \frac{1}{n}\left|\{Y^n_1, \ldots, Y^n_n\} \cap \left(\mathcal{T}^n_{\text{left}}(u_i), \mathcal{T}^n_{\text{right}}(u_i)\right)\right| + o_\mathbb{P}(1).$$

Introducing the empirical distribution function of the $(Y^n_i)_{i \leq n}$

$$F_n(y) := \frac{1}{n}\left|\{Y^n_1, \ldots, Y^n_n\} \cap (-\infty, y)\right|,$$

we have

$$t(\mathcal{T}^n, u_i) = F_n\left(\mathcal{T}^n_{\text{right}}(u_i)\right) - F_n\left(\mathcal{T}^n_{\text{left}}(u_i)\right) + o_\mathbb{P}(1).$$

For each fixed n, $(Y^n_i)_{1 \leq i \leq n}$ are i.i.d. random variables in $[0, 1]$. Since μ is a permuton, it satisfies (1), and the common distribution of the Y^n_i's is the uniform distribution. From Proposition 1.5, we infer that F_n converges a.s. uniformly on $[0, 1]$ to the identity function (the earlier use of Skorohod's representation theorem implies that the $(Y^n_i)_{1 \leq i \leq n}$ are coupled in a nontrivial way for different values of n, but Proposition 1.5 applies nevertheless).

Moreover, the above discussion implies that $\mathcal{T}^n_{\text{right}}(u_i)$ and $\mathcal{T}^n_{\text{left}}(u_i)$ converge a.s. to $\mathcal{T}^\infty_{\text{right}}(u_i)$ and $\mathcal{T}^\infty_{\text{left}}(u_i)$ respectively. Therefore, a.s. in the probability space created by the application of Skorohod's representation theorem, we have that, for all $i \leq d$,

$$t(\mathcal{T}^n, u_i) = \mathcal{T}^\infty_{\text{right}}(u_i) - \mathcal{T}^\infty_{\text{left}}(u_i) + o_\mathbb{P}(1) = \psi_{\mu_0}(u_i) + o_\mathbb{P}(1).$$

Since a.s. (joint) convergence implies (joint) convergence in distribution, (11) is proved, concluding the proof of Theorem 1.2. □

3. SOME COMPARISON ARGUMENTS AND CONSEQUENCES

3.1. **Height modification by adding/removing points.** Given a tree $\mathcal{T}$, a *chain* in $\mathcal{T}$ is a subset C of its nodes such that for every pair (v, w) in C, either v is an ancestor of w, or the converse. We note that the height of $\mathcal{T}$ is simply the maximal size of a chain in $\mathcal{T}$, minus 1. We extend this definition to point sets as follows: given a set of points $\mathcal{P} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with distinct coordinates, we say that $C \subseteq \mathcal{P}$ is a chain in $\mathcal{T}\langle\mathcal{P}\rangle$ if the corresponding nodes in $\mathcal{T}\langle\mathcal{P}\rangle$ form a chain.

Lemma 3.1. *Let $y = (y_1, \ldots, y_n)$ be a list of distinct numbers and $\mathcal{T} = \mathcal{T}\langle y\rangle$ be the associated BST. If $i < j$ are two indices then the following are equivalent:*

(i) *y_i is an ancestor of y_j in $\mathcal{T}$ (the converse cannot hold);*
(ii) *there is no $k < i$ such that y_k is between y_i and y_j, i.e. such that $(y_i - y_k)(y_j - y_k) < 0$.*

Proof. As in Section 1.5, we see $\mathcal{T}$ as a top tree, formed by the insertion of the first $i - 1$ elements, and hanging trees. The condition (ii) above exactly stipulates that y_i and y_j are in the same hanging tree in this decomposition. If this is the case, then y_i is the first vertex inserted in this hanging tree, and therefore is its root, so that y_i is indeed an ancestor of y_j. Conversely, if y_i and y_j are in different hanging trees, then y_i cannot be an ancestor of y_j. □

Lemma 3.2. *Let $\mathcal{P}_- \subseteq \mathcal{P}_+$ be two sets of points with distinct x- and y-coordinates. Then, for any chain C of $\mathcal{T}\langle\mathcal{P}_+\rangle$, the set $C \cap \mathcal{P}_-$ is a chain of $\mathcal{T}\langle\mathcal{P}_-\rangle$. Consequently, if C is a chain of maximal size in $\mathcal{T}\langle\mathcal{P}_+\rangle$, we have*

$$h\left(\mathcal{T}\langle\mathcal{P}_-\rangle\right) \geq h\left(\mathcal{T}\langle\mathcal{P}_+\rangle\right) - \left|C \cap (\mathcal{P}_+ \setminus \mathcal{P}_-)\right|.$$

Proof. Write $y^+ = (y_1^+, y_2^+, \dots)$ and $y^- = (y_1^-, y_2^-, \dots)$ for the sequences of y-coordinates of $\mathcal{P}_+$ and $\mathcal{P}_-$ read by increasing x-coordinate, so that $\mathcal{T}\langle \mathcal{P}_+ \rangle = \mathcal{T}\langle y^+ \rangle$ and $\mathcal{T}\langle \mathcal{P}_- \rangle = \mathcal{T}\langle y^- \rangle$. Using Lemma 3.1, if $C = (y_{i_1}^+, \dots, y_{i_\ell}^+)$ is a chain of $\mathcal{T}\langle y^+ \rangle$ then for any triple of indices $k < i_j < i_{j'}$ one has

$$\left(y_{i_j}^+ - y_k^+ \right) \left(y_{i_{j'}}^+ - y_k^+ \right) > 0.$$

Since $y^- \subseteq y^+$, this property still holds when restricted to y^-, and therefore $C \cap y^-$ is a chain of $\mathcal{T}\langle y^- \rangle$.

Now if $\mathcal{C}$ is a chain of maximal size in $\mathcal{T}\langle y^+ \rangle$, one has $h\left(\mathcal{T}\langle y^+ \rangle\right) = |\mathcal{C}| - 1$. But $\mathcal{C} \cap y^-$ is a chain in $\mathcal{T}\langle y^- \rangle$, implying

$$h\left(\mathcal{T}\langle y^- \rangle\right) \geq \left| \mathcal{C} \cap y^- \right| - 1 = |\mathcal{C}| - \left| \mathcal{C} \cap (y^+ \setminus y^-) \right| - 1 = h\left(\mathcal{T}\langle y^+ \rangle\right) - \left| \mathcal{C} \cap (y^+ \setminus y^-) \right|. \qquad \square$$

Combining the above lemma with standard thinning properties of Poisson point processes, we get the following useful proposition.

Proposition 3.3. *Let $\rho_- \leq \rho_+$ be two intensity functions defined on the same support $S \subseteq \mathbb{R}^2$, and $\mathcal{P}_-, \mathcal{P}_+$ be two Poisson point processes with intensities ρ_- and ρ_+. Then, we have*

$$h\left(\mathcal{T}\langle \mathcal{P}_- \rangle\right) \succeq \text{BINOMIAL}\left(1 + h\left(\mathcal{T}\langle \mathcal{P}_+ \rangle\right), \inf_{(x,y) \in S} \frac{\rho_-(x,y)}{\rho_+(x,y)}\right) - 1.$$

Proof. Write $r = \inf_{(x,y) \in S} \frac{\rho_-(x,y)}{\rho_+(x,y)}$ where, by convention, $\frac{\rho_-(x,y)}{\rho_+(x,y)} = 1$ if $\rho_+(x,y) = 0$. We couple $\mathcal{P}_+$ and $\mathcal{P}_-$ according to the classical thinning process, meaning that $\mathcal{P}_-$ is constructed by keeping each point (x,y) of $\mathcal{P}_+$ independently with probability $\rho_-(x,y)/\rho_+(x,y) \geq r$. Let C be a set of points of $\mathcal{P}_+$ corresponding to a chain of maximum length in $\mathcal{T}\langle \mathcal{P}_+ \rangle$ and denote by $K = |C \cap \mathcal{P}_-| = |C| - \left| C \cap (\mathcal{P}_+ \setminus \mathcal{P}_-) \right|$ the number of points on this chain kept by the thinning procedure. Using the thinning process, conditionally given $\mathcal{P}_+$, the following stochastic dominance holds:

$$K \succeq \text{BINOMIAL}\left(|C|, r\right).$$

Then by Lemma 3.2, since $|C| = h\left(\mathcal{T}\langle \mathcal{P}_+ \rangle\right) + 1$, we have

$$h\left(\mathcal{T}\langle \mathcal{P}_- \rangle\right) \geq h\left(\mathcal{T}\langle \mathcal{P}_+ \rangle\right) - \left| C \cap (\mathcal{P}_+ \setminus \mathcal{P}_-) \right| = K - 1. \qquad \square$$

Remark 3. Considering as in Lemma 3.2 two sets of points $\mathcal{P}_- \subseteq \mathcal{P}_+$ with distinct x- and y-coordinates, the height $h\left(\mathcal{T}\langle \mathcal{P}_- \rangle\right)$ can be much bigger than $h\left(\mathcal{T}\langle \mathcal{P}_+ \rangle\right)$, even while removing a single point. To see this, take n odd, let $\mathcal{P}_- = \{(i/n, i/n), 0 < i < n\}$, and $\mathcal{P}_+ = \mathcal{P}_- \cup \{(0, 1/2)\}$. The points in $\mathcal{P}_-$ form an increasing sequence, so that $\mathcal{T}\langle \mathcal{P}_- \rangle$ consists in a single branch growing to the right, and has height $n-2$. On the other hand, the root in $\mathcal{T}\langle \mathcal{P}_+ \rangle$ has label $1/2$ and divides the tree into two equal parts of size $(n-1)/2$. Each of this part has height $(n-1)/2 - 1$, so that $h\left(\mathcal{T}\langle \mathcal{P}_+ \rangle\right) = (n-1)/2$.

3.2. **A de-Poissonization result.** The comparison lemma can also be used to de-Poissonize convergence results for the height of BSTs. Recall that $\mathcal{P}_\mu^n$ and $\mathcal{P}_\mu^N$ denote respectively a sample of n i.i.d. random points with law μ, and a Poisson point process with intensity $n\mu$.

Theorem 3.4. *Let μ be a permuton. Let $f : \mathbb{N} \to [1, \infty)$ be a function such that there exists $\frac{1}{2} < \alpha < 1$ satisfying*

$$(13) \qquad \sup_{|\delta| \leq n^\alpha} \left| \frac{f(n+\delta)}{f(n)} - 1 \right| \xrightarrow[n \to \infty]{} 0.$$

Then we have

$$\frac{h\left(\mathcal{T}\langle \mathcal{P}_\mu^n \rangle\right)}{f(n)} \xrightarrow{\mathbb{P}} 1 \iff \frac{h\left(\mathcal{T}\langle \mathcal{P}_\mu^N \rangle\right)}{f(n)} \xrightarrow{\mathbb{P}} 1$$

as $n \to \infty$. Moreover all powers of $\frac{h\left(\mathcal{T}\langle \mathcal{P}_\mu^n \rangle\right)}{f(n)}$ are uniformly integrable if and only if all powers of $\frac{h\left(\mathcal{T}\langle \mathcal{P}_\mu^N \rangle\right)}{f(n)}$ are uniformly integrable.

Proof. First suppose that $\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle)}{f(n)} \xrightarrow{\mathbb{P}} 1$. Then $\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle)}{f(N)} \xrightarrow{\mathbb{P}} 1$ as $n \to \infty$. Moreover, since $\alpha > 1/2$, w.h.p. it holds that $n - n^\alpha \leq N \leq n + n^\alpha$, so the regularity hypothesis (13) on f implies $f(n)/f(N) \xrightarrow{\mathbb{P}} 1$ and we conclude that $\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle)}{f(n)} \xrightarrow{\mathbb{P}} 1$.

Now suppose that all powers of $\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle)}{f(n)}$ are uniformly integrable, or equivalently that for every integer p, the sequence $\mathbb{E}\left[\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle)^p}{f(n)^p}\right]$ is bounded in n. This implies that the sequence $\mathbb{E}\left[\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle)^p}{f(N)^p}\right]$ is also bounded in n. Therefore:

$$\mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle\right)^p}{f(n)^p}\right] \leq \mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle\right)^p}{f(n)^p}\mathbf{1}_{|N-n|\leq n^\alpha}\right] + \mathbb{E}\left[N^p\mathbf{1}_{|N-n|>n^\alpha}\right]$$

$$\leq \mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle\right)^p}{f(N)^p}\right]\sup_{|\delta|\leq n^\alpha}\frac{f(n+\delta)^p}{f(n)^p} + \sqrt{\mathbb{E}\left[N^{2p}\right]\mathbb{P}\left(|N-n|>n^\alpha\right)}.$$

The first term is bounded by assumption, while the second one is easily proved to tend to 0, using Lemma 1.4 and Equation (3). We conclude that, for any p, the quantity $\mathbb{E}\left[\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle)^p}{f(n)^p}\right]$ is bounded in n, and thus, all powers of $\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle)}{f(n)}$ are uniformly integrable.

The converse implications require de-Poissonization and are more subtle. The global idea is to bound the height of $\mathcal{T}\langle\mathcal{P}_\mu^n\rangle$ by two Poissonized versions, using Lemma 3.2. However, making this idea work (both for probability convergence and for L^p boundedness) requires some computation.

Let $N_+ \sim \text{POISSON}(n + n^\alpha)$ and $\mathcal{P}^{N_+}$ be a Poisson point process of intensity $(n+n^\alpha)\mu$. Since $\alpha > \frac{1}{2}$, the event $E_+ = \{N_+ \geq n\}$ holds w.h.p. as $n \to \infty$. Under E_+, define $\mathcal{P}^n$ as a uniform subset of size n in $\mathcal{P}^{N_+}$. Then $\mathcal{P}^n$ is a set of n i.i.d. points distributed under μ. Similarly, let $N_- \sim \text{POISSON}(n - n^\alpha)$ and note that the event $E_- = \{N_- \leq n\}$ holds w.h.p. as $n \to \infty$. Under the event E_- and conditionally given N_-, define $\mathcal{P}^{N_-}$ as a uniform subset of size N_- in $\mathcal{P}^n$. Then conditionally under E_-, the set $\mathcal{P}^{N_-}$ is distributed like a Poisson point process of intensity $(n - n^\alpha)\mu$ conditioned to have at most n points. By applying Lemma 3.2 to both $\mathcal{P}^n \subseteq \mathcal{P}^{N_+}$ and $\mathcal{P}^{N_-} \subseteq \mathcal{P}^n$, conditionally under $E = E_- \cap E_+$, we have that

$$(14) \qquad h\left(\mathcal{T}\langle\mathcal{P}^{N_+}\rangle\right) - \left|C_+ \cap \left(\mathcal{P}^{N_+} \setminus \mathcal{P}^n\right)\right| \leq h\left(\mathcal{T}\langle\mathcal{P}^n\rangle\right) \leq h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right) + \left|C \cap \left(\mathcal{P}^n \setminus \mathcal{P}^{N_-}\right)\right|$$

where C_+, C are arbitrary chains of maximal size in $\mathcal{T}\langle\mathcal{P}^{N_+}\rangle$ and $\mathcal{T}\langle\mathcal{P}^n\rangle$ respectively. Under E and conditionally on N_+ and $h\left(\mathcal{T}\langle\mathcal{P}^{N_+}\rangle\right)$, we may apply Lemma 1.3 to obtain:

$$(15) \qquad \left|C_+ \cap \left(\mathcal{P}^{N_+} \setminus \mathcal{P}^n\right)\right| \preceq \text{BINOMIAL}\left(1 + h\left(\mathcal{T}\langle\mathcal{P}^{N_+}\rangle\right), \frac{N_+ - n}{N_+ - h\left(\mathcal{T}\langle\mathcal{P}^{N_+}\rangle\right) - 1}\right).$$

Similarly, assuming E and conditionally given N_- and $h\left(\mathcal{T}\langle\mathcal{P}^n\rangle\right)$, it holds that:

$$(16) \qquad \left|C \cap \left(\mathcal{P}^n \setminus \mathcal{P}^{N_-}\right)\right| \preceq \text{BINOMIAL}\left(1 + h\left(\mathcal{T}\langle\mathcal{P}^n\rangle\right), \frac{n - N_-}{n - h\left(\mathcal{T}\langle\mathcal{P}^n\rangle\right) - 1}\right).$$

We now suppose that $\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle)}{f(n)} \xrightarrow{\mathbb{P}} 1$ as n goes to infinity, and we want to prove that $\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle)}{f(n)}$ tends to 1 in probability as well. Using the regularity hypothesis (13) on f, we have $\frac{h(\mathcal{T}\langle\mathcal{P}^{N_\pm}\rangle)}{f(n)} \xrightarrow{\mathbb{P}} 1$. Throughout the proof, we fix constants $\beta \in (\alpha, 1)$ and $\delta > 0$.

Lower bound on $h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right)$. Using (14), as n goes to infinity:

$$\mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right) \leq (1-\delta)f(n)\right]$$
$$\leq \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right) \leq (1-\delta)f(n) \wedge E\right] + \mathbb{P}[E^c]$$
$$\leq \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}^{N_+}\rangle\right) - \left|C_+ \cap (\mathcal{P}^{N_+}\setminus\mathcal{P}^n)\right| \leq (1-\delta)f(n)\right] + o(1)$$
$$\leq \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}^{N_+}\rangle\right) \leq (1-\delta/2)f(n)\right] + \mathbb{P}\left[\left|C_+ \cap (\mathcal{P}^{N_+}\setminus\mathcal{P}^n)\right| \geq \frac{\delta}{2}f(n)\right] + o(1).$$

In the last line, we can conclude directly from the convergence $\frac{h\left(\mathcal{T}\langle\mathcal{P}^{N_+}\rangle\right)}{f(n)} \xrightarrow{\mathbb{P}} 1$ that the first probability is a $o(1)$, but the second one requires more attention.

Notice that under the event $\left|C_+ \cap (\mathcal{P}^{N_+}\setminus\mathcal{P}^n)\right| \geq \frac{\delta}{2}f(n)$, we have $f(n) \leq \frac{2}{\delta}(N_+ - n)$. Moreover, w.h.p. as $n \to \infty$, it holds that $N_+ - n \leq n^\beta$ and $1 + h\left(\mathcal{T}\langle\mathcal{P}_\mu^{N_+}\rangle\right) \leq 2f(n)$. Thus the parameters of the binomial random variable in (15) are bounded w.h.p. by $2f(n)$ and $2n^{\beta-1}$ respectively (recall that $\beta < 1$). Denote by S_n a BINOMIAL $\left(\lfloor 2f(n)\rfloor, 2n^{\beta-1}\right)$ random variable to obtain the following:

$$\mathbb{P}\left[\left|C_+ \cap (\mathcal{P}^{N_+}\setminus\mathcal{P}^n)\right| \geq \frac{\delta}{2}f(n)\right] \leq \mathbb{P}\left[S_n \geq \frac{\delta}{2}f(n)\right] + o(1) \leq \frac{2\,\mathbb{E}[S_n]}{\delta f(n)} + o(1) = o(1)$$

This concludes the proof that

$$\mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right) \leq (1-\delta)f(n)\right] \xrightarrow[n\to\infty]{} 0.$$

Upper bound on $h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right)$. For each $n \in \mathbb{N}$ we distinguish between two cases.

- Suppose $f(n) \geq n^\beta$. Then we use (14) as before:

$$\mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right) \geq (1+\delta)f(n)\right]$$
$$\leq \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right) + \left|C \cap (\mathcal{P}^n\setminus\mathcal{P}^{N_-})\right| \geq (1+\delta)f(n)\right] + o(1)$$
$$\leq \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right) \geq (1+\delta/2)f(n)\right] + \mathbb{P}\left[n - N_- \geq \delta n^\beta/2\right] + o(1)$$
$$= o(1).$$

- Suppose $f(n) \leq n^\beta$. From (14) and the trivial bound $\left|C \cap (\mathcal{P}^n\setminus\mathcal{P}^{N_-})\right| \leq n - N_-$, we have $n - h\left(\mathcal{T}\langle\mathcal{P}^n\rangle\right) \geq N_- - h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)$. But, w.h.p., $N_- \geq n - n^\beta$ and $h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right) \leq 2f(n) \leq 2n^\beta$, implying that $n - h\left(\mathcal{T}\langle\mathcal{P}^n\rangle\right) \geq n - 3n^\beta$. Using again that $n - N^- \leq n^\beta$ w.h.p., the probability parameter of the binomial random variable in (16) tends to 0 in probability. Hence w.h.p., one has $\left|C \cap (\mathcal{P}^n\setminus\mathcal{P}^{N_-})\right| \leq \frac{\delta/2}{1+\delta}h\left(\mathcal{T}\langle\mathcal{P}^n\rangle\right)$ (recall that δ is an arbitrary positive constant, and thus, so is $\frac{\delta/2}{1+\delta}$). The upper bound of (14) yields w.h.p.:

$$\left(1 - \frac{\delta/2}{1+\delta}\right)h\left(\mathcal{T}\langle\mathcal{P}^n\rangle\right) \leq h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right),$$

and then

$$\mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right) \geq (1+\delta)f(n)\right]$$
$$= \mathbb{P}\left[\left(1 - \frac{\delta/2}{1+\delta}\right)h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right) \geq (1+\delta/2)f(n)\right]$$
$$\leq \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right) \geq (1+\delta/2)f(n)\right] + o(1)$$
$$= o(1).$$

We have thus proved that $\frac{h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right)}{f(n)} \xrightarrow{\mathbb{P}} 1$.

<u>Uniform integrability.</u> Finally suppose that for every integer p, the sequence $\mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}^N_\mu\rangle\right)^p}{f(n)^p}\right]$ is bounded in n. Thanks to our hypothesis on f, the sequence $\mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)^p}{f(n)^p}\mathbf{1}_E\right]$ is also bounded in n. Then (14), together with the convexity inequality $(a+b)^p \leq 2^{p-1}(a^p+b^p)$ for $a,b \geq 0$ and $p \geq 1$, yields:

$$(17) \quad \mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}^n_\mu\rangle\right)^p}{f(n)^p}\right] \leq n^p \cdot \mathbb{P}(E^c) + 2^{p-1}\mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)^p}{f(n)^p}\mathbf{1}_E\right] + 2^{p-1}\mathbb{E}\left[\frac{\left|C \cap (\mathcal{P}^n \setminus \mathcal{P}^{N_-})\right|^p}{f(n)^p}\mathbf{1}_E\right].$$

But $n^p \cdot \mathbb{P}(E^c)$ converges to 0 (as a consequence of Lemma 1.4, the probability actually decreases at least as fast as $e^{-n^{2\alpha-1}}$), while the second term was already identified as being bounded in n. Let us consider the last term, which we split as follows

$$(18) \quad \mathbb{E}\left[\frac{\left|C \cap (\mathcal{P}^n \setminus \mathcal{P}^{N_-})\right|^p}{f(n)^p}\mathbf{1}_E\right]$$
$$= \mathbb{E}\left[\frac{\left|C \cap (\mathcal{P}^n \setminus \mathcal{P}^{N_-})\right|^p}{f(n)^p}\mathbf{1}_E\mathbf{1}_{h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)\geq n^\beta}\right] + \mathbb{E}\left[\frac{\left|C \cap (\mathcal{P}^n \setminus \mathcal{P}^{N_-})\right|^p}{f(n)^p}\mathbf{1}_E\mathbf{1}_{h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)< n^\beta}\right]$$

For the first term, we bound $\left|C \cap (\mathcal{P}^n \setminus \mathcal{P}^{N_-})\right|$ by $(n-N_-)$ and, using the Cauchy–Schwarz inequality, we get

$$(19) \quad \mathbb{E}\left[\frac{\left|C \cap (\mathcal{P}^n \setminus \mathcal{P}^{N_-})\right|^p}{f(n)^p}\mathbf{1}_E\mathbf{1}_{h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)\geq n^\beta}\right] \leq \mathbb{E}\left[\frac{(n-N_-)^p}{n^{\beta p}}\frac{h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)^p}{f(n)^p}\mathbf{1}_E\right]$$
$$\leq \sqrt{\mathbb{E}\left[\frac{(n-N_-)^{2p}}{n^{2\beta p}}\right]\mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)^{2p}}{f(n)^{2p}}\mathbf{1}_E\right]}.$$

Since $n-N_- = n^\alpha - (N_- - \mathbb{E}[N_-])$ and since it is well known that $\frac{1}{\sqrt{n-n^\alpha}}(N_- - \mathbb{E}[N_-])$ converges in distribution and in moments to a standard Gaussian random variable, the first expectation under the square root tends to 0 as n tends to ∞ (recall that $\beta > \alpha > \frac{1}{2}$). Also, it has been observed above that the second expectation under the square root is bounded in n. Thus the left-hand side of (19) tends to 0 as n tends to ∞.

We now consider the second term in (18). From Lemma 1.4, one has $N_- \geq n - n^\beta$ outside a set of exponentially small probability. When this holds together with the events E and $h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right) < n^\beta$, one has, for n large enough,

$$\frac{n-N_-}{n-h\left(\mathcal{T}\langle\mathcal{P}^n\rangle\right)-1} \leq \frac{n^\beta}{n-n^\beta-1} \leq 2n^{\beta-1}.$$

Hence, using (16), and writing S_n for a BINOMIAL $\left(1+h\left(\mathcal{T}\langle\mathcal{P}^n_\mu\rangle\right), 2n^{\beta-1}\right)$ random variable, we get:

$$\mathbb{E}\left[\frac{\left|C \cap (\mathcal{P}^n \setminus \mathcal{P}^{N_-})\right|^p}{f(n)^p}\mathbf{1}_E\mathbf{1}_{h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)< n^\beta}\right] \leq \mathbb{E}\left[\frac{S_n^p}{f(n)^p}\right] + n^p\,\mathbb{P}[N_- < n - n^\beta] = \mathbb{E}\left[\frac{S_n^p}{f(n)^p}\right] + o(1).$$

The moments of a BINOMIAL (M, q) random variable X are easily bounded by $\mathbb{E}[X^p] \leq M^p q$, implying that

$$\mathbb{E}\left[\frac{\left|C \cap (\mathcal{P}^n \setminus \mathcal{P}^{N_-})\right|^p}{f(n)^p}\mathbf{1}_E\mathbf{1}_{h\left(\mathcal{T}\langle\mathcal{P}^{N_-}\rangle\right)< n^\beta}\right] \leq \mathbb{E}\left[\frac{\left(1+h\left(\mathcal{T}\langle\mathcal{P}^n_\mu\rangle\right)\right)^p}{f(n)^p}\right]2n^{\beta-1} + o(1).$$

Using this last estimate, (17)-(18) and the fact that (19) tends to 0, we get that

$$\mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}^n_\mu\rangle\right)^p}{f(n)^p}\right] \leq \mathcal{O}(1) + 2n^{\beta-1}\mathbb{E}\left[\frac{\left(1+h\left(\mathcal{T}\langle\mathcal{P}^n_\mu\rangle\right)\right)^p}{f(n)^p}\right].$$

Using again that $(a+b)^p \leq 2^{p-1}(a^p+b^p)$ and since $2n^{\beta-1}$ tends to 0, this implies that $\mathbb{E}\left[\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle)^p}{f(n)^p}\right]$

is bounded (for all p), proving the uniform integrability of all powers of $\frac{h(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle)}{f(n)}$. $\square$

3.3. **A connection with monotone subsequences and extreme deviation bounds.** We start by recalling standard definitions. Let σ be a permutation of $\{1,\ldots,n\}$. An increasing subsequence of σ is a sequence of indices $i_1 < \cdots < i_k$ such that $\sigma(i_1) < \cdots < \sigma(i_k)$. The maximum length of an increasing subsequence of σ is then denoted by $\mathrm{LIS}(\sigma)$. We define similarly $\mathrm{LDS}(\sigma)$, the maximum length of a decreasing subsequence of σ.

Lemma 3.5. *Let σ be a permutation of $\{1,\ldots,n\}$. Then*

$$h(\mathcal{T}\langle\sigma\rangle) \leq \mathrm{LIS}(\sigma) + \mathrm{LDS}(\sigma).$$

Proof. Let $i_1 < \cdots < i_k$ be a sequence of integers such that $\sigma(i_1),\ldots,\sigma(i_k)$ label nodes on a chain C of $\mathcal{T}\langle\sigma\rangle$. Define $\mathcal{I}_\mathcal{R}$ (resp. $\mathcal{I}_\mathcal{L}$) as the family of i_j's such that the node following $\sigma(i_j)$ in C lies in its right subtree (resp. left subtree). By construction of the BST, $\mathcal{I}_\mathcal{R} \cup \{i_k\}$ and $\mathcal{I}_\mathcal{L} \cup \{i_k\}$ form respectively an increasing and a decreasing subsequence of σ. The lemma follows. $\square$

Combining this lemma with [BGS22, Proposition 3.2], we get that for any integrable function ρ, the quantity $\frac{1}{n}h(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle)$ tends to 0 in probability, as $n \to \infty$. We will need a more quantitative version of this, valid only for bounded functions ρ. We start with an extreme deviation bound[1] for the longest monotone subsequences.

Lemma 3.6. *For each integer n, let σ^n be a uniform permutation of $\{1,\ldots,n\}$. Then we have*

$$\mathbb{P}\left[\mathrm{LIS}(\sigma^n) \geq \frac{n}{\log n}\right] \leq \exp(-n + o(n)).$$

Proof. This is a straightforward application of the first moment method. Write, with $k = \lfloor\frac{n}{\log n}\rfloor$:

$$\mathbb{P}\left[\mathrm{LIS}(\sigma^n) \geq \frac{n}{\log n}\right] \leq \mathbb{E}[\text{number of increasing subsequences of length } k \text{ in } \sigma^n] = \frac{1}{k!}\binom{n}{k}$$

Now, using Stirling's formula along with $k = \lfloor\frac{n}{\log n}\rfloor = o(n)$, we obtain

$$\frac{1}{k!}\binom{n}{k} = \frac{n!}{k!^2(n-k)!} = e^{n\log n - n - 2k\log k - (n-k)\log(n-k) + (n-k) + o(n)} = e^{-n+o(n)}.   \square$$

Corollary 3.7. *For any $M > 0$ and $\varepsilon > 0$, there exists $n_0 = n_0(M,\varepsilon)$ such that the following holds. For any $0 < \zeta \leq 1$, any function $\rho : [0,1]^2 \to [0,\infty)$ bounded by M and supported on some rectangle $[a,b] \times [c,d]$ with $(b-a)(d-c) \leq \zeta$, and for any integer $n > n_0/\zeta$:*

$$\mathbb{P}\left[h(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle) > 2\varepsilon\zeta n\right] \leq 4\exp\left(-\frac{\varepsilon}{2}\zeta n \log(\zeta n)\right).$$

Proof. By increasing the size of the rectangle $[a,b] \times [c,d]$, we can assume that $(b-a)(d-c) = \zeta$. Now let us stretch ρ from $[a,b] \times [c,d]$ onto $[0,1]^2$. Define

$$g : (x,y) \in [a,b] \times [c,d] \mapsto \left(\frac{x-a}{b-a}, \frac{y-c}{d-c}\right) \in [0,1]^2.$$

Then $\widetilde{\mathcal{P}} = g(\mathcal{P}_\rho^N)$ is a Poisson point process with intensity $n\zeta\rho \circ g^{-1}$ on $[0,1]^2$. Moreover, the transformation g does not change the relative order of points, so $\mathcal{T}\langle\widetilde{\mathcal{P}}\rangle$ and $\mathcal{T}\langle\mathcal{P}_\rho^N\rangle$ have the same shape. From now on we work with $\widetilde{\mathcal{P}}$.

---

[1]We use the term "extreme devitation" since, for the longest increasing subsequence, the usual "large deviation framework" consists in studying $\mathbb{P}[\mathrm{LIS}(\sigma^n) \geq x\sqrt{n}]$ for $x > 2$, see [Sep98]. We look here at much rarer events.

The density $\rho \circ g^{-1}$ is bounded above by M on $[0,1]^2$. Thus, by a standard thickening procedure, we can construct a Poisson point process $\widehat{\mathcal{P}}$ with constant intensity $n\zeta M$ such that a.s. $\widetilde{\mathcal{P}} \subset \widehat{\mathcal{P}}$. Let $\widetilde{\sigma}$ and $\widehat{\sigma}$ be the permutations induced by $\widetilde{\mathcal{P}}$ and $\widehat{\mathcal{P}}$, respectively. Using Lemma 3.5, a.s. it holds that:

$$h\left(\mathcal{T}\langle\widetilde{\mathcal{P}}\rangle\right) \leq \mathrm{LIS}\left(\widetilde{\sigma}\right) + \mathrm{LDS}\left(\widetilde{\sigma}\right) \leq \mathrm{LIS}\left(\widehat{\sigma}\right) + \mathrm{LDS}\left(\widehat{\sigma}\right). \tag{20}$$

The permutation $\widehat{\sigma}$ has a random size $\widehat{N}$, which follows a POISSON $(n\zeta M)$ law. Moreover, conditionally to its size, it is uniformly distributed. By Lemma 1.4, for any n large enough such that $\varepsilon \log(\zeta n) \geq M$, we get:

$$\mathbb{P}\left[\widehat{N} > \varepsilon\zeta n \log(\zeta n)\right] \leq e^{-n\zeta M} \left(\frac{eM}{\varepsilon \log(\zeta n)}\right)^{\varepsilon\zeta n \log(\zeta n)} \leq \exp\left(\varepsilon\zeta n \log(\zeta n) \log\left(\frac{eM}{\varepsilon \log(\zeta n)}\right)\right).$$

For large enough ζn (with a threshold depending on M and ε), it holds that $\log\left(\frac{eM}{\varepsilon \log(\zeta n)}\right) \leq -1$, and thus

$$\mathbb{P}\left[\widehat{N} > \varepsilon\zeta n \log(\zeta n)\right] \leq \exp\left(-\varepsilon\zeta n \log(\zeta n)\right) \tag{21}$$

Define $n' = \lfloor \varepsilon\zeta n \log(\zeta n) \rfloor$ and write $\sigma^{n'}$ for a uniform permutation of $\{1, \ldots, n'\}$. As $\frac{n'}{\log(n')} \leq \varepsilon\zeta n$ for large enough ζn, we have

$$\mathbb{P}\left[\mathrm{LIS}\left(\widehat{\sigma}\right) > \varepsilon\zeta n\right] \leq \mathbb{P}\left[\mathrm{LIS}\left(\sigma^{n'}\right) > \frac{n'}{\log(n')}\right] + \mathbb{P}\left[\widehat{N} > \varepsilon\zeta n \log(\zeta n)\right].$$

Using Lemma 3.6 and Equation (21), we get that, for large enough ζn,

$$\mathbb{P}\left[\mathrm{LIS}\left(\widehat{\sigma}\right) > \varepsilon\zeta n\right] \leq e^{-n' + o(n')} + e^{-\varepsilon\zeta n \log(\zeta n)} \leq 2e^{-\frac{1}{2}\varepsilon\zeta n \log(\zeta n)}.$$

The same holds for $\mathrm{LDS}\left(\widehat{\sigma}\right)$, and Equation (20) allows us to conclude the proof of the corollary. □

4. HEIGHT OF BSTs OF PERMUTON SAMPLES

Before starting the proof, let us introduce some notation related to the decomposition presented in the introduction (Section 1.5). Let $\mathcal{P} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ be a set of points in $(0,1)^2$ with distinct x- and y-coordinates. For any $\beta \in (0,1)$, write $\mathcal{P}(\beta) = \mathcal{P} \cap ([0, \beta] \times [0,1])$ for the set of points in a band of width β on the left. Now write $K_\beta = |\mathcal{P}(\beta)|$ and let $y_{(1)} < \cdots < y_{(K_\beta)}$ be the ordered y-coordinates of the points in $\mathcal{P}(\beta)$. Then for each integer $0 \leq k \leq K_\beta$, define $I_k = (y_{(k)}, y_{(k+1)})$ where we used the convention $y_{(0)} = 0$ and $y_{(K_\beta+1)} = 1$. In words, $I_0, \ldots, I_{K_\beta}$ are the gaps between the points $\{0, y_{(1)}, \ldots, y_{(K_\beta)}, 1\}$, enumerated from lowest to highest. Finally, define

$$\mathcal{P}_k(\beta) := \mathcal{P} \cap \left((\beta, 1] \times I_k\right).$$

Recall that $\mathcal{T}\langle\mathcal{P}(\beta)\rangle$ and $\left(\mathcal{T}\langle\mathcal{P}_k(\beta)\rangle\right)_{0 \leq k \leq K_\beta}$ are respectively called the *top tree* and the *hanging trees* of $\mathcal{T}\langle\mathcal{P}\rangle$. One can see that the top and hanging trees are indeed subtrees of $\mathcal{T}\langle\mathcal{P}(\beta)\rangle$. The entire tree can then be reconstructed as follows: start with $\mathcal{T}\langle\mathcal{P}(\beta)\rangle$, and for each $0 < k < K_\beta$ do the following. Write v_k, resp. v_{k+1}, for the node labeled $y_{(k)}$, resp. $y_{(k+1)}$, in $\mathcal{T}\langle\mathcal{P}(\beta)\rangle$ and notice that necessarily one is an ancestor of the other. If v_k is deeper than v_{k+1} then attach $\mathcal{T}\langle\mathcal{P}_k(\beta)\rangle$ to the right of v_k, otherwise attach it to the left of v_{k+1}. Finally, attach $\mathcal{T}\langle\mathcal{P}_0(\beta)\rangle$ to the left of the node labeled $y_{(1)}$ and $\mathcal{T}\langle\mathcal{P}_{K_\beta}(\beta)\rangle$ to the right of the node labeled $y_{(K_\beta)}$. The reader can go back to Figure 5 for an illustration. This construction yields the following lemma:

Lemma 4.1. *Let $\mathcal{P}$ be any set of points of $[0,1]^2$ with distinct x- and distinct y-coordinates. Then for any $\beta \in (0,1)$:*

$$h\left(\mathcal{T}\langle\mathcal{P}(\beta)\rangle\right) \leq h\left(\mathcal{T}\langle\mathcal{P}\rangle\right) \leq h\left(\mathcal{T}\langle\mathcal{P}(\beta)\rangle\right) + 1 + \max_{0 \leq k \leq K_\beta}\left\{h\left(\mathcal{T}\langle\mathcal{P}_k(\beta)\rangle\right)\right\}.$$

4.1. **Controlling the height of the top tree.**

Proposition 4.2. *Let $R = [x_1, x_2] \times [y_1, y_2]$ be a rectangle with non-empty interior and $\rho : R \to (0, \infty)$ be a continuous, positive intensity function. For each integer n, let $\mathcal{P}_\rho^N$ be a Poisson point process with intensity $n\rho$. Let $m \leq M$ be positive real numbers such that $m \leq \rho \leq M$ holds a.e. on R and write*

$$\eta = \frac{M - m}{m}\,.$$

Then for any $\varepsilon > 0$, we have

$$\lim_{n \to \infty} \mathbb{P}\left[\left|\frac{h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)}{c^* \log n} - 1\right| > \eta + \varepsilon\right] = 0\,.$$

Moreover, for any $p > 0$, the sequence of random variables $\frac{h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)^p}{\log(n)^p}$ is uniformly integrable.

Proof. Write $\zeta = (x_2 - x_1)(y_2 - y_1) > 0$ for the area of R. Also note that

$$\frac{m}{M} = 1 + \frac{m - M}{M} \geq 1 - \frac{M - m}{m} = 1 - \eta \qquad ; \qquad \frac{M}{m} = 1 + \frac{M - m}{m} = 1 + \eta\,.$$

Using Proposition 3.3 with $\rho_- = n\rho$ and $\rho_+ = nM$ on R, we obtain

$$h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right) \succeq \textsc{Binomial}\left(1 + h\left(\mathcal{T}\langle\mathcal{P}_+\rangle\right), \frac{m}{M}\right) - 1,$$

where $\mathcal{P}_+ := \mathcal{P}_{\rho_+}^N$. Since ρ_+ is a constant density, the tree $\mathcal{T}\langle\mathcal{P}_+\rangle$ is the BST of a uniform permutation of random size $\textsc{Poisson}(n\zeta M)$. According to [Dev86, Theorem 5.1], $h\left(\mathcal{T}\langle\mathcal{P}_+\rangle\right)$ then behaves as $c^* \log(|\mathcal{P}_+|)$ as $n \to \infty$ in probability, which leads to

$$1 + h\left(\mathcal{T}\langle\mathcal{P}_+\rangle\right) = c^* \log(n\zeta M) + o_\mathbb{P}(\log n) = c^* \log n + o_\mathbb{P}(\log n)\,.$$

Using that a $\textsc{Binomial}(a \log n, m/M)$ random variable is concentrated around its mean $(am/M) \log n$, we get:

$$h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right) \geq \frac{m}{M}\left(c^* \log n - o_\mathbb{P}(\log n)\right) \geq \left(1 - \eta - o_\mathbb{P}(1)\right)c^* \log n\,.$$

Similarly, using Proposition 3.3 with $\rho_- = nm$ and $\rho_+ = n\rho$ we obtain

(22) $$h\left(\mathcal{T}\langle\mathcal{P}_-\rangle\right) \succeq \textsc{Binomial}\left(1 + h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right), \frac{m}{M}\right) - 1,$$

where $\mathcal{P}_- := \mathcal{P}_{\rho_-}^N$. As before, we observe that $\mathcal{T}\langle\mathcal{P}_-\rangle$ is the BST of a uniform permutation of random size $\textsc{Poisson}(n\zeta m)$. This implies

$$h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right) \leq \frac{M}{m}c^* \log(n\zeta m) + o_\mathbb{P}(\log n) = \left(1 + \eta + o_\mathbb{P}(1)\right)c^* \log n\,,$$

which concludes the proof of the first claim.

For the uniform integrability claim, let us fix $p > 0$ and establish boundedness of $\mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)^p}{\log(n)^p}\right]$ in n. Conditionally given $h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)$, write S_n for a $\textsc{Binomial}\left(1 + h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right), \frac{m}{M}\right)$ random variable substracted by 1. Then, using Hoeffding's inequality:

$$\mathbb{P}\left[S_n < \frac{m}{2M}\left(1 + h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)\right) - 1 \;\bigg|\; h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)\right] \leq e^{-\frac{m^2}{2M^2}\left(1 + h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)\right)}$$

and therefore, for any $n \geq e$:

$$
\mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)^p}{\log(n)^p}\right]
$$

$$
\leq \mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)^p}{\log(n)^p}\mathbf{1}_{S_n<\frac{m}{2M}\left(1+h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)\right)-1}\right] + \mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)^p}{\log(n)^p}\mathbf{1}_{S_n\geq\frac{m}{2M}\left(1+h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)\right)-1}\right]
$$

$$
\leq \mathbb{E}\left[h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)^p e^{-\frac{m^2}{2M^2}\left(1+h\left(\mathcal{T}\langle\mathcal{P}_\rho^N\rangle\right)\right)}\right] + \mathbb{E}\left[\frac{\left((2M/m)\cdot(S_n+1)-1\right)^p}{\log(n)^p}\right].
$$

Since the function $x \mapsto x^p e^{-\frac{m^2}{2M^2}(1+x)}$ is bounded over $\mathbb{R}_+$, the first term is bounded in n. As for the second term, we use (22) along with $(a+b)^p \leq 2^{p-1}(a^p+b^p)$ to deduce:

$$
\mathbb{E}\left[\frac{\left((2M/m)\cdot(S_n+1)-1\right)^p}{\log(n)^p}\right] \leq \left(\frac{2M}{m}\right)^p \mathbb{E}\left[\frac{(S_n+1)^p}{\log(n)^p}\right]
$$

$$
\leq 2^{2p-1}\left(\frac{M}{m}\right)^p\left(\mathbb{E}\left[\frac{h\left(\mathcal{T}\langle\mathcal{P}_-\rangle\right)^p}{\log(n)^p}\right] + \frac{1}{\log(n)^p}\right)
$$

which is bounded in n by [Dev86, Lemma 3.1] and Poisson estimates. This concludes the proof of the uniform integrability claim. □

The weakness of the previous proposition is that η, which depends on the rectangle under consideration, might be big. In the next statement we show that, for continuous positive densities ρ, it is possible to choose rectangles for which the corresponding η is small.

Corollary 4.3. *Let D be a compact domain in the plane and $\rho: D \to (0,\infty)$ be a continuous, positive intensity function. Then for any $\varepsilon > 0$, there exists $\beta > 0$ such that for any $\beta' \leq \beta$ and any rectangle $R = [x_1, x_1+\beta'] \times [y_1, y_2]$ with non-empty interior contained in D:*

$$
\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{h\left(\mathcal{T}\langle\mathcal{P}_\rho^N \cap R\rangle\right)}{c^*\log n} - 1\right| > \varepsilon\right] = 0.
$$

In particular, taking $x_1 = y_1 = 0$ and $y_2 = 1$, the tree $\mathcal{T}\langle\mathcal{P}_\rho^N \cap R\rangle$ is the top tree $\mathcal{T}\langle\mathcal{P}_\rho^N(\beta)\rangle$ defined at the beginning of Section 4.

Proof. Let $\varepsilon > 0$ and assume that $\varepsilon < \min_D \rho$. By uniform continuity of ρ, find $\beta > 0$ such that for any $(x,y),(x',y') \in D$:

$$
|x-x'| + |y-y'| \leq \beta \quad\Longrightarrow\quad |\rho(x,y) - \rho(x',y')| \leq \varepsilon.
$$

Then fix $\beta' \leq \beta$ and consider any rectangle $R = [x_1, x_1+\beta'] \times [y_1, y_2]$ contained in D. Define

$$
f : y \in [y_1, y_2] \mapsto \int_{y_1}^y \rho(x_1, t)dt \quad\text{and}\quad g : y \in [y_1, y_2] \mapsto y_1 + (y_2 - y_1)f(y)/f(y_2).
$$

The function g is a $\mathcal{C}^1$ increasing map from $[y_1, y_2]$ onto itself. Let $\widetilde{\mathcal{P}}$ denote the set of points obtained after applying the transformation $(x,y) \mapsto (x, g(y))$ to $\mathcal{P}_\rho^N \cap R$. This transformation does not change the relative orders of points, therefore $\mathcal{T}\langle\widetilde{\mathcal{P}}\rangle$ and $\mathcal{T}\langle\mathcal{P}_\rho^N \cap R\rangle$ have the same shape. Additionally, $\widetilde{\mathcal{P}}$ follows the law of a Poisson point process with intensity

$$
n\frac{\rho(x, g^{-1}(y))}{g'(g^{-1}(y))} = n\frac{f(y_2)}{y_2 - y_1}\frac{\rho(x, g^{-1}(y))}{\rho(x_1, g^{-1}(y))}
$$

on R. Thus we can apply Proposition 4.2 with

$$
m = \frac{f(y_2)}{y_2 - y_1}\left(1 - \frac{\varepsilon}{\min_D \rho}\right), \quad M = \frac{f(y_2)}{y_2 - y_1}\left(1 + \frac{\varepsilon}{\min_D \rho}\right), \quad \eta = \frac{2\varepsilon}{\min_D \rho - \varepsilon},
$$

to obtain:

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{h\left(\mathcal{T}\langle\tilde{\mathcal{P}}\rangle\right)}{c^* \log n} - 1\right| > \eta + \varepsilon\right] = 0\,.$$

Since this holds for any small enough $\varepsilon > 0$ and since η goes to 0 when ε goes to 0, the result follows. $\qquad\square$

4.2. **Some bounds on the hanging trees.** As above, let $\mathcal{P}^N$ be a Poisson point process on $[0,1]^2$ with intensity $n\mu$. We use the notations at the beginning of Section 4. Moreover, for each $k \leq |\mathcal{P}^N(\beta)|$, we let $\zeta_k = |I_k|$ be the size of the k-th gap, and $\mathcal{P}_k^N(\beta)$ be the points of $\mathcal{P}^N$ in the horizontal band $(\beta, 1] \times I_k$. The sizes of the bands and the number of points in each band are then controlled by the following proposition.

Proposition 4.4. *Let μ be a permuton. Assume that there exists $\beta > 0$ such that $\mu_{/[0,\beta]\times[0,1]}$ has a continuous and positive density $\rho : [0,\beta] \times [0,1] \to (0,\infty)$. Then the following holds.*

(1) *There exists α such that*

$$\lim_{n\to+\infty} \mathbb{P}\left[\max_k \zeta_k > \alpha \frac{\log n}{n}\right] = 0.$$

(2) *The sequence of random variables*

$$\left(\frac{1}{\log n} \max_{0\leq k \leq |\mathcal{P}(\beta)|} |\mathcal{P}_k^N(\beta)|\right)^p$$

is uniformly integrable for any $p \geq 1$.

Proof. Let $\ell = \lceil n/(a \log n)\rceil$, where the constant $a \geq 1$ will be specified later, and let us divide vertically the rectangles $[0,\beta] \times [0,1]$ and $(\beta,1] \times [0,1]$, each into ℓ cells of the same size. Namely we set, for $0 \leq i < \ell$,

$$C_i = [0,\beta] \times \left[\tfrac{i}{\ell}, \tfrac{i+1}{\ell}\right], \quad D_i = (\beta,1] \times \left[\tfrac{i}{\ell}, \tfrac{i+1}{\ell}\right].$$

For each i, the probability that $\mathcal{P}^N \cap C_i$ is empty equals $e^{-n\mu(C_i)} \leq e^{-n\frac{\beta m}{\ell}}$, where m is a lower bound for the density ρ on $[0,\beta] \times [0,1]$. Call E the event that all C_i contain at least one point of $\mathcal{P}^N$. It follows from the above computation along with a union bound that

$$(23) \qquad \mathbb{P}[E^c] = \mathbb{P}\Big[\exists i : \mathcal{P}^N \cap C_i = \emptyset\Big] \leq \ell\, e^{-n\frac{\beta m}{\ell}} \leq \left\lceil\frac{n}{a\log n}\right\rceil e^{-\beta m a \log n + o(\log n)}$$

as $n \to \infty$. When E is satisfied, for all k it holds that $\zeta_k \leq \frac{2}{\ell} \leq \frac{2a\log n}{n}$, implying

$$\mathbb{P}\left[\max_k \zeta_k > \frac{2a\log n}{n}\right] \leq \mathbb{P}\big(E^c\big) \leq \left\lceil\frac{n}{a\log n}\right\rceil e^{-\beta m a \log n + o(\log n)}.$$

The latter bound tends to 0 if we choose $a > 1/(\beta m)$, showing item (1).

For item (2) we observe that, assuming E, any given band $(\beta,1] \times I_k$ intersects at most two cells D_i, implying

$$\max_{0\leq k \leq |\mathcal{P}(\beta)|} |\mathcal{P}_k^N(\beta)| \leq 2 \max_{0\leq i \leq \ell-1} |\mathcal{P}^N \cap D_i|.$$

The random variables $|\mathcal{P}^N \cap D_i|$ are POISSON $(n\mu(D_i))$ distributed. Since $\mu(D_i) \leq \mu([0,1]\times[\tfrac{i}{\ell},\tfrac{i+1}{\ell}]) = \frac{1}{\ell} \leq \frac{a\log n}{n}$, this implies the following estimates for $b > 0$:

$$\mathbb{P}\left[\max_{0\leq k \leq |\mathcal{P}(\beta)|} |\mathcal{P}_k^N(\beta)| \geq 2b\log n\,,\, E\right] \leq \mathbb{P}\left[\max_{0\leq i \leq \ell-1} |\mathcal{P}^N \cap D_i| \geq b\log n\right]$$

$$\leq \ell\, \mathbb{P}\Big[\text{POISSON}\,(a\log n) \geq b\log n\Big].$$

Using Lemma 1.4, for $b > a$ we deduce

$$(24) \qquad \mathbb{P}\left[\max_{0\leq k \leq |\mathcal{P}(\beta)|} |\mathcal{P}_k^N(\beta)| \geq 2b\log n\,,\, E\right] \leq \left(\frac{ea}{b}\right)^{b\log n}\left\lceil\frac{n}{a\log n}\right\rceil e^{-a\log n}.$$

For $a \geq 1$ and $b \geq ea$ the right-hand side is uniformly bounded by $(ea/b)^b$ for any $n \geq e$.

We now have all the necessary estimates to conclude the proof of item (2). Recall that we want to prove uniform integrability of all powers of $\frac{1}{\log n} \max_{0 \leq k \leq |\mathcal{P}(\beta)|} |\mathcal{P}_k^N(\beta)|$, which is equivalent to boundedness in n of all its moments. Fix $p > 0$. We have, using the layer cake representation and (23) and (3):

$$(25) \quad \mathbb{E}\left[\left(\frac{\max_{0 \leq k \leq |\mathcal{P}(\beta)|} |\mathcal{P}_k^N(\beta)|}{\log n}\right)^p\right] \leq \mathbb{E}\left[|\mathcal{P}^N|^p\right]\mathbb{P}[E^c] + \mathbb{E}\left[\left(\frac{\max_{0 \leq k \leq |\mathcal{P}(\beta)|} |\mathcal{P}_k^N(\beta)|}{\log n}\right)^p \mathbf{1}_E\right]$$

$$\leq n^p(1 + o(1))\left\lceil\frac{n}{a \log n}\right\rceil e^{-\beta m a \log n + o(\log n)} + p\int_0^\infty s^{p-1}\mathbb{P}\left[\frac{\max_{0 \leq k \leq |\mathcal{P}(\beta)|} |\mathcal{P}_k^N(\beta)|}{\log n} \geq s \wedge E\right]ds.$$

We choose $a > \frac{p+1}{\beta m}$, so that the first term tends to 0, and thus, is a bounded sequence in n. The second term is bounded as follows, using (24) for $s \geq 2ea$ and simply bounding the probability by 1 otherwise:

$$\int_0^\infty s^{p-1}\mathbb{P}\left[\frac{\max_{0 \leq k \leq |\mathcal{P}(\beta)|} |\mathcal{P}_k^N(\beta)|}{\log n} \geq s \wedge E\right]ds \leq \int_0^\infty s^{p-1}\min\left(1, \left(\frac{2ea}{s}\right)^{s/2}\right)ds.$$

This integral is finite and independent of n, and we conclude that (25) is bounded in n. The proposition is proved. □

Item (1) can be further used to control the maximal height of a hanging tree in $T\langle\mathcal{P}_\mu^N\rangle$.

Proposition 4.5. *Let μ be a permuton satisfying* (A1), *i.e. μ has an upper bounded density ρ on $[0,1]$, which is positive and continuous on $[0,\beta] \times [0,1]$ for some $\beta > 0$. Then we have the following convergence in probability as n goes to infinity:*

$$\frac{1}{\log n} \max_{0 \leq k \leq |\mathcal{P}(\beta)|}\left\{h\left(\mathcal{T}\langle\mathcal{P}_k^N(\beta)\rangle\right)\right\} \longrightarrow 0.$$

Proof. From Proposition 4.4, item (1), there exists $\alpha > 0$ such that $\max_k \zeta_k < \alpha\frac{\log n}{n}$ w.h.p. as $n \to \infty$. We shall work conditionally given $\mathcal{P}^N(\beta)$, and assume that $\max_k \zeta_k < \alpha\frac{\log n}{n}$. In particular, as in Section 4, we let $\{y_{(1)} < \cdots < y_{(K_\beta)}\}$ be the ordered y-coordinates of the points in $\mathcal{P}(\beta)$, with $K_\beta = |\mathcal{P}(\beta)|$, and set by convention $y_{(0)} = 0$ and $y_{(K_\beta+1)} = 1$. Then for each $0 \leq k \leq |\mathcal{P}(\beta)|$, the family $\mathcal{P}_k^N(\beta)$ is distributed like a Poisson point process with intensity $n\rho_{/[\beta,1] \times [y_{(k)}, y_{(k+1)}]}$.

For any $k \leq |\mathcal{P}(\beta)|$, since $\zeta_k < \alpha\frac{\log n}{n}$, Corollary 3.7 applies with ρ restricted to $[\beta, 1] \times [y_{(k)}, y_{(k+1)}]$ and $\zeta = \alpha\frac{\log n}{n}$. Thus for $n\zeta = \alpha\log n$ large enough:

$$\mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_k^N(\beta)\rangle\right) > \varepsilon\log n\right] = \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_k^N(\beta)\rangle\right) > 2\frac{\varepsilon}{2\alpha}\zeta n\right] \leq 4\exp\left[-\frac{\varepsilon}{4\alpha}(\alpha\log n)\log(\alpha\log n)\right].$$

A union bound then implies that, still conditionally given $\mathcal{P}(\beta)$ and assuming $\max_k \zeta_k < \alpha\frac{\log n}{n}$, one has

$$\mathbb{P}\left[\frac{1}{\log n}\max_{0 \leq k \leq |\mathcal{P}(\beta)|} h\left(\mathcal{T}\langle\mathcal{P}_k^N(\beta)\rangle\right) > \varepsilon\right] \leq (|\mathcal{P}(\beta)| + 1) \cdot 4\exp\left[-\frac{\varepsilon}{4\alpha}(\alpha\log n)\log(\alpha\log n)\right].$$

But w.h.p., the inequality $\max_k \zeta_k < \alpha\frac{\log n}{n}$ indeed holds and $|\mathcal{P}(\beta)| < n$, so the unconditioned probability tends to 0 as n tends to infinity:

$$\mathbb{P}\left[\frac{1}{\log n}\max_{0 \leq k \leq |\mathcal{P}(\beta)|} h\left(\mathcal{T}\langle\mathcal{P}_k^N(\beta)\rangle\right) > \varepsilon\right] \xrightarrow[n\to\infty]{} 0.$$

This holds for any $\varepsilon > 0$, proving the proposition. □

Remark 4. Item (1) in Proposition 4.4 could alternatively have been derived using standard results on the maximal gap, also called *maximal spacing*, between i.i.d. uniform random variables; see e.g. [Slu78].

Indeed, by applying a thinning procedure, $\max_k \zeta_k$ is bounded above by the maximal gap between Poisson $(n\beta m)$ i.i.d. uniform variables in $[0,1]$, which is known to concentrate around $\log(n)/(n\beta m)$.

4.3. Concluding the proof of the height theorem.

First, we can deduce uniform integrability of all powers of $\frac{h(\mathcal{T}\langle \mathcal{P}^N\rangle)}{\log n}$, under a hypothesis which is slightly weaker than (A1) (more precisely, we only make an assumption regarding the behavior on $(\beta,1] \times [0,1]$).

Proposition 4.6. *Let μ be a permuton. Assume that there exists $\beta > 0$ such that $\mu_{/[0,\beta]\times[0,1]}$ has a continuous and positive density $\rho : [0,\beta] \times [0,1] \to (0,\infty)$. Then, for any $p > 0$, the family of random variables*

$$\left(\left(\frac{h\left(\mathcal{T}\langle \mathcal{P}_\mu^N\rangle\right)}{\log n} \right)^p \right)_{n \geq 2}$$

is uniformly integrable.

Proof. This follows immediately from Lemma 4.1, together with Propositions 4.2 and 4.4, using the trivial bound $h\left(\mathcal{T}\langle \mathcal{P}_k^N(\beta)\rangle\right) \leq \left|\mathcal{P}_k^N(\beta)\right|$ for the hanging trees. $\qquad\square$

Now we can combine our results to establish Theorem 1.1 under assumption (A1).

Proof of Theorem 1.1. Thanks to Theorem 3.4, it suffices to prove the theorem in its Poissonized version. We shall work with $\mathcal{P}^N$, a Poisson point process of intensity $n\mu$.

Fix $\varepsilon > 0$. Let D be a compact neighborhood of $\{0\} \times [0,1]$ on which ρ is continuous and positive, and let $\beta = \beta(\varepsilon) > 0$ be given by Corollary 4.3 applied to ρ on D. Therefore

$$\lim_{n\to\infty} \mathbb{P}\left[\left| \frac{h\left(\mathcal{T}\langle \mathcal{P}^N(\beta)\rangle\right)}{c^* \log n} - 1 \right| > \varepsilon \right] = 0$$

where $\mathcal{T}\langle \mathcal{P}^N(\beta)\rangle = \mathcal{T}\langle \mathcal{P}^N \cap ([0,\beta] \times [0,1])\rangle$ is the top tree of $\mathcal{T}\langle \mathcal{P}^N\rangle$. Furthermore, by Proposition 4.5, the quantity

$$\frac{1}{\log n} \max_{0 \leq k \leq |\mathcal{P}(\beta)|} \left\{ h\left(\mathcal{T}\langle \mathcal{P}_k(\beta)\rangle\right) \right\}$$

converges in probability to 0. Combining this with Lemma 4.1, we conclude that $\frac{1}{\log n} h\left(\mathcal{T}\langle \mathcal{P}^N\rangle\right)$ converges in probability to c^*. Finally, Proposition 4.6 implies uniform integrability of all powers, and thus L^p convergence for all $p \geq 1$. $\qquad\square$

5. Extra results

5.1. Permutons with partially vanishing densities on the left edge and binary search trees of polynomial depth.

Let $E := \left\{(x,y) \in [0,1]^2 : x \leq y \leq x + \frac{1}{2} \text{ or } y \leq x - \frac{1}{2}\right\}$, see Figure 6 (left). It is straight-forward to check that the measure μ defined by $\mu(A) := 2\,\text{Leb}(A \cap E)$ has uniform marginals, i.e. is a permuton.
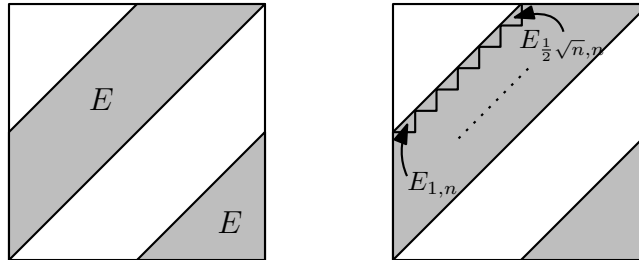

FIGURE 6. Left: the support E of the permuton of Proposition 5.1. Right: the sets $E_{i,n}$ involved in the proof.

Proposition 5.1. *Let μ be the above permuton. Then, for any $\varepsilon > 0$:*

$$\lim_{n\to\infty} \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right) \geq \tfrac{1}{2}(1 - \tfrac{1}{e} - \varepsilon)\sqrt{n}\right] = 1.$$

Proof. We first consider a Poisson point process $\mathcal{P}_\mu^N$ of intensity $n\mu$. We say that a point (x, y) in $\mathcal{P}_\mu^N$ is a record if there is no point in $([0, x) \times (y, 1]) \cap \mathcal{P}_\mu^N$, i.e. if there is no point of $\mathcal{P}_\mu^N$ above and to the left of (x, y). It is easily seen that in the construction of $\mathcal{T}\langle\mathcal{P}_\mu^N\rangle$, a point is inserted in the right-most branch of the tree if and only if it is a record. Hence the number of nodes on that right-most branch is the number of records in $\mathcal{P}_\mu^N$, which we will denote by $\text{rec}(\mathcal{P}_\mu^N)$. Therefore, $h\left(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle\right) \geq \text{rec}(\mathcal{P}_\mu^N) - 1$, and we will prove a lower bound in probability for $\text{rec}(\mathcal{P}_\mu^N)$.

For any positive integer $i \leq \frac{1}{2}\sqrt{n}$, let us define $x_i := i/\sqrt{n}$ and

$$E_{i,n} := \{(x, y) : x_{i-1} \leq y - \tfrac{1}{2} \leq x \leq x_i\} \subset E,$$

as shown on Figure 6. We have $\mu(E_{i,n}) = 2\,\text{Leb}(E_{i,n}) = 1/n$, implying that each $E_{i,n}$ contains POISSON (1) points in $\mathcal{P}_\mu^N$. Moreover, these numbers are independent random variables. Hence, for any $\varepsilon > 0$, w.h.p., at least a fraction $(1 - \frac{1}{e} - \varepsilon)$ of the sets $E_{i,n}$ (for $i \leq \frac{1}{2}\sqrt{n}$) contain a point in $\mathcal{P}_\mu^N$. Each non-empty $E_{i,n}$ contains at least one record, implying $\text{rec}(\mathcal{P}_\mu^N) \geq \frac{1}{2}(1 - \frac{1}{e} - \varepsilon)\sqrt{n}$ w.h.p. This shows

$$\lim_{n\to\infty} \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle\right) \geq \tfrac{1}{2}(1 - \tfrac{1}{e} - \varepsilon)\sqrt{n}\right] = 1.$$

The same result for $\mathcal{P}_\mu^n$ is deduced using the de-Poissonization techniques of Theorem 3.4. $\qquad\square$

On the other hand, for any permuton μ with a bounded density ρ, we have $h\left(\mathcal{T}\langle\mathcal{P}_\mu^n\rangle\right) = \mathcal{O}(\sqrt{n})$ with high probability. Indeed, this follows from Lemma 3.5 and the proof method of [Dub23b, Proposition 1.3]. Hence the lower bound given in Proposition 5.1 is optimal up to a multiplicative constant.

The above example can be modified to a permuton with a continuous density. Let $\varphi : [0, 1/4] \to \mathbb{R}$ be a continuous function satisfying $\varphi(0) = 0$ and $\int_0^{1/4} \varphi(t)dt = 1/2$. We then define

$$\rho(x, y) = \varphi\left(\text{dist}_{L_1}\left((x, y), E^c \cup \{(1, 0)\}\right)\right).$$

Going cyclically along any horizontal line, the L_1 distance above grows linearly from 0 to 1/4 and then decreases linearly again from 1/4 to 0. The same holds along vertical lines. Since $\int_0^{1/4} \varphi(t)dt = 1/2$, this implies that the measure $\mu = \rho(x, y)dxdy$ is a permuton supported on E. Moreover with the notation of the above proof, we have

$$\mu(E_{i,n}) = \int_0^{1/\sqrt{n}} (1/\sqrt{n} - t)\varphi(t)dt.$$

Choosing e.g. $\varphi(t) \sim t^\delta$ for small t, we have

$$\mu(E_{i,n}) \sim \frac{n^{-1-\delta/2}}{(\delta + 1)(\delta + 2)},$$

implying that each $E_{i,n}$ contains a point in $\mathcal{P}_\mu^N$ with probability roughly $n^{-\delta/2}$. Hence, w.h.p. at least $\Theta(n^{(1-\delta)/2})$ sets among the $E_{i,n}$ are non-empty, showing that the height $h\left(\mathcal{T}\langle\mathcal{P}_\mu^N\rangle\right)$ has polynomial growth. This illustrates the importance of the positivity assumption on the density near the left edge of the square made in Theorem 1.1.

5.2. **Permutons with positive densities on a band and binary search trees of large logarithmic depth.** For $\beta > 0$, we consider the permuton μ_β which has a mass β uniformly distributed on the band $[0, \beta] \times [0, 1]$ and a mass $1 - \beta$ uniformly distributed on the line segment from $(\beta, 0)$ to $(1, 1)$. These permutons satisfy the following.

Proposition 5.2. *For any $\beta > 0$ and $\varepsilon > 0$, we have*

$$(26) \qquad \lim_{n\to+\infty} \mathbb{P}\left[h\left(\mathcal{T}\langle\mathcal{P}_{\mu_\beta}^n\rangle\right) \geq \frac{1-\beta}{\beta+\varepsilon}\log n\right] = 1.$$

Proof. As usual we first consider a Poissonized version, namely let $\mathcal{P}_{\mu_\beta}^N$ be a Poisson point process of intensity $n\mu_\beta$. As in Section 4, we let $y_{(1)} < \cdots < y_{(K_\beta)}$ be the ordered y-coordinates of the points in $\mathcal{P}(\beta)$. The number K_β is distributed as a Poisson(βn) random variable and thus $K_\beta/(\beta n)$ converges to 1 in probability. Finally, we set $\zeta_* = \max_k y_{(k+1)} - y_{(k)}$ with the convention $y_{(0)} = 0$ and $y_{(K_\beta+1)} = 1$. Conditionally given K_β, this is the maximal gap (or maximal spacing) defined by K_β i.i.d. uniform random variables in $[0,1]$, and, from a result of Slud [Slu78], the quotient $\frac{\zeta_* K_\beta}{\log(K_\beta)}$ converges to 1 in probability. Hence $\frac{\zeta_* \beta n}{\log(n)}$ converges to 1 in probability.

We now observe that, since the points of $\mathcal{P}_{\mu_\beta}^N$ with x-coordinates bigger than β are in increasing order, all hanging trees $\mathcal{T}\langle \mathcal{P}_k(\beta)\rangle$ consist of a single right branch, and thus $h\left(\mathcal{T}\langle \mathcal{P}_k(\beta)\rangle\right) = |\mathcal{P}_k(\beta)| - 1$ for any $k \leq K_\beta$. Conditionally given $\mathcal{P}(\beta)$, and letting k_0 be such that $y_{(k_0+1)} - y_{(k_0)} = \zeta_*$, the number of elements in $\mathcal{P}_{k_0}(\beta)$ concentrates around $n(1-\beta)\zeta_*$, which is close to $\frac{1-\beta}{\beta}\log n$ in probability. Since $h\left(\mathcal{T}\langle \mathcal{P}_{\mu_\beta}^N\rangle\right) \geq h\left(\mathcal{T}\langle \mathcal{P}_{k_0}(\beta)\rangle\right)$, this proved (26) with $\mathcal{P}_{\mu_\beta}^n$ replaced by its Poissonized version $\mathcal{P}_{\mu_\beta}^N$. The proposition follows using the de-Poissonization techniques of Theorem 3.4. $\qquad\square$

Our bound is once again optimal up to a multiplicative constant. Indeed, the permutons μ_β satisfy the hypotheses of Proposition 4.6, which proves that all powers of $\frac{h\left(\mathcal{T}\langle \mathcal{P}_{\mu_\beta}^n\rangle\right)}{\log n}$ are uniformly integrable.

5.3. **A lower bound result.** As last result in this paper, we emphasize that the lower bound in Theorem 1.1 holds under a rather weak hypothesis. This can be seen as a partial result towards Conjecture 1.

Proposition 5.3. *Let μ be a permuton. Suppose there exists $0 \leq y \leq 1$ such that μ admits a continuous, positive density ρ on a neighborhood of the point $(0, y)$. Then for any $\varepsilon > 0$:*

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{h\left(\mathcal{T}\langle \mathcal{P}_\mu^N\rangle\right)}{c^* \log n} \leq 1 - \varepsilon\right] = 0.$$

Proof. This is relatively easy to prove, based on some intermediate results, established while proving Theorem 1.1. Let D be a compact neighborhood of the point $(0, y)$ on which μ admits a continuous, positive density ρ. Fix $\varepsilon > 0$ and let $\beta = \beta(\varepsilon) > 0$ be given by Corollary 4.3. We can take $0 < \delta \leq \beta$ so that there exists at least one rectangle $R_0 = [0, \delta] \times [y_0 - \delta, y_0 + \delta]$ contained in D. Then by Corollary 4.3:

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{h\left(\mathcal{T}\langle \mathcal{P}_\mu^N \cap R_0\rangle\right)}{c^* \log n} \leq 1 - \varepsilon\right] = 0.$$

Moreover, using Lemma 3.1, one can see that each chain of $\mathcal{T}\langle \mathcal{P}_\mu^N \cap R_0\rangle$ is still a chain in $\mathcal{T}\langle \mathcal{P}_\mu^N\rangle$. Therefore $h\left(\mathcal{T}\langle \mathcal{P}_\mu^N \cap R_0\rangle\right) \leq h\left(\mathcal{T}\langle \mathcal{P}_\mu^N\rangle\right)$ a.s., and this concludes the proof. $\qquad\square$

6. Acknowledgements

References

[ABC21]   L. Addario-Berry and B. Corsini. The height of Mallows trees. *Ann. Probab.*, 49(5):2220–2271, 2021.

[BBF+20]   F. Bassino, M. Bouvel, V. Féray, L. Gerin, M. Maazoun, and A. Pierrot. Universal limits of substitution-closed permutation classes. *J. Eur. Math. Soc. (JEMS)*, 22(11):3565–3639, 2020.

[BDMW23]   J. Borga, S. Das, S. Mukherjee, and P. Winkler. Large deviation principle for random permutations. *International Mathematics Research Notices*, rnad096, 2023.

[BGS22]   J. Borga, E. Gwynne, and X. Sun. Permutons, meanders, and sle-decorated liouville quantum gravity, 2022. Preprint arXiv:2207.02319.

[Bil99]   P. Billingsley. *Convergence of probability measures.* Wiley Ser. Probab. Stat. Chichester: Wiley, 2nd ed. edition, 1999.

[Bil12]   P. Billingsley. *Probability and measure. Anniversary edition.* Hoboken, NJ: John Wiley & Sons, 2012.

[Cor23]   B. Corsini. The height of record-biased trees. *Random Structures & Algorithms*, 62(3):623–644, 2023.

[Dev86]   L. Devroye. A note on the height of binary search trees. *Journal of the ACM (JACM)*, 33(3):489–498, 1986.

[Dub23a]   V. Dubach. Increasing subsequences of linear size in random permutations and the Robinson–Schensted tableaux of permutons, 2023. Preprint arXiv:2307.05768.

[Dub23b]   V. Dubach. Locally uniform random permutations with large increasing subsequences. *Combinatorial Theory*, 3(3), 2023.

[DZ95]   J.-D. Deuschel and O. Zeitouni. Limiting curves for i.i.d. records. *Ann. Probab.*, 23(2):852–878, 1995.

[Grü23]   R. Grübel. A note on limits of sequences of binary trees. *Discrete Mathematics & Theoretical Computer Science*, 25(Analysis of Algorithms), 2023.

[HKM$^+$13] C. Hoppen, Y. Kohayakawa, C. G. Moreira, B. Ráth, and R. M. Sampaio. Limits of permutation sequences. *J. Comb. Theory, Ser. B*, 103(1):93–113, 2013.

[MS13]   C. Mueller and S. Starr. The length of the longest increasing subsequence of a random Mallows permutation. *J. Theor. Probab.*, 26(2):514–540, 2013.

[Pit84]   B. Pittel. On growing random binary trees. *J. Math. Anal. Appl.*, 103:461–480, 1984.

[Sep98]   T. Seppäläinen. Large deviations for increasing sequences on the plane. *Probab. Theory Relat. Fields*, 112(2):221–244, 1998.

[Sjö23]   J. Sjöstrand. Monotone subsequences in locally uniform random permutations. *Ann. Probab.*, 51(4):1502–1547, 2023.

[Slu78]   E. Slud. Entropy and maximal spacings for random partitions. *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, 41:341–352, 1978.

[Sta09]   S. Starr. Thermodynamic limit for the Mallows model on S_n. *J. Math. Phys.*, 50(9):095208, 15, 2009.

(BC) Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands

Email address: benoitcorsini@gmail.com

(VD,VF) Université de Lorraine, CNRS, IECL, F-54000 Nancy, France

Email address: victor.dubach@univ-lorraine.fr, valentin.feray@univ-lorraine.fr