On the statistical properties of the isolation forest anomaly detection method

Bruno Pelletier

On the statistical properties of the isolation forest anomaly detection method

Bruno Pelletier

Département de Mathématiques, IRMAR – UMR CNRS 6625, Université Rennes II, France

July 15, 2024

Abstract

Isolation forest is a popular method for anomaly detection introduced in Liu et al. (2008, 2012). Nonetheless, its statistical properties are little understood. We study the scoring function that is induced by the isolation forest over a finite sample at the limit when the number of trees tends to infinity, based on an analytical expression that we derive. The isolation forest method is proved to be effective at detecting geometrically isolated points within a finite sample. We then study the large sample limit of the scoring function in random designs as well as in sequences of regular designs and we find that the isolation forest method performs as a detector of the support of the underlying distribution. We also find that dense clustered anomalies are not detected asymptotically by the isolation forest method, a phenomenon known as the masking effect, but that isolation forest anomaly detection is robust to training with normal data sparsely contaminated by anomalies. Numerical examples are provided that confirm the theoretical results.

Keywords—Anomaly detection, Isolation forest, Binary tree, Scoring, Recursive partition.

1 Introduction

Anomaly detection can be defined as the task of identifying those observations that stand out in some way or another. Precise definition of an anomaly may vary, and multiple proposals have been formulated in the literature, some quoted in Samariya and Thakkar (2023), but they all share a concept of deviation from a reference situation. See also Foorthuis (2021), who assembled a comprehensive typology of data anomalies. The subject of anomaly detection has a long history and is an important line of research in statistics and machine learning. Many methods and algorithms have been introduced so far and applied succesfully in a wide range of domains, such as intrusion detection in network systems, fraud detection in credit card, insurance and finance, fault detection in complex systems, and medical monitoring, to name a few; see for instance the reviews given in Chandola et al. (2009) and more recently in Samariya and Thakkar (2023). Anomaly detection is also closely related to outlier detection (Aggarwal, 2017; Barnett & Lewis, 1994) and novelty detection (Markou & Singh, 2003a, 2003b). In fact, albeit referring to somewhat distinct purposes, these terms are often used exchangeably and the detection methods typically apply to either context (Chandola et al., 2009).

Liu et al. (2008, 2012) introduced the isolation forest method for anomaly dectection. The underlying idea is that anomalies, and even small clusters of anomalies, are somehow more susceptible of being isolated than non-abnormal data, so that it might be easier to isolate them from the rest of the data than it is for non-abnormal data (Liu et al., 2010). In essence, the isolation forest method consists in building an ensemble of random binary space partitioning trees, called *isolation trees*. When an isolation tree is fully grown so that each element of the

partition induced by the tree contains exactly one data point, the data points are isolated one from the other at the leaves (external nodes) of the tree (assuming that such a data separation is possible). The path length from the root of an isolation tree to one of its external node gives the number of recursive partitioning steps that are needed to isolate the data point located at that node. Then under the premise that anomalies are easier to isolate than the other data, anomalies are expected to correspond to shorter path lengths in the isolation trees than for non-abnormal data. An *anomaly score* is then produced for each data point by aggregating the respective path lengths over the ensemble of isolation trees.

In the algorithm of Liu et al. (2008, 2012) and in practice, the isolation trees are constructed on subsamples and are subjected to a maximum height limit. In this case, the leaves of the trees may contain more than one data point. The subsample size and the maximum height count among the tuning parameters of the algorithm, together with the number of trees. Based on the anomaly scores returned by the isolation forest algorithm, observations can then be labeled as being abnormal or not by simple thresholding. The scoring strategy proposed in Liu et al. (2008) produces anomaly scores normalized between 0 and 1, with a score close to 1 indicating evidence of an anomaly and a score close to 0 indicating confidence in a non-abnormal data. It is worth noticing that, by the very nature of an isolation tree of being a partitioning tree, the anomaly scores that are produced by the forest at the data points readily extend to a *scoring function* defined on the whole sample space. This offers the possibity of scoring new, unseen data, based on the forest built on the training data. An example use case of this mode of operation is in the design of an anomaly-based intrusion detection system in which the training data is known to represent the normal state of the system and, as such, does not contain any anomaly (see for instance Khraisat et al., 2019).

Since its introduction, the isolation forest procedure has gained strong popularity and several variants and extensions have then been proposed. In the original algorithm of Liu et al. (2008), the isolation trees are grown recursively by operating random binary splits along the coordinate axes. The axis at which the split occurs is selected uniformly at random, and next the value of the split is selected uniformly at random over the range of the projected data onto that axis. Partitioning along the axes may induce artifacts in the shape of the scoring function, as reported in Hariri et al. (2021), who then proposed a variant called the extended isolation forest, which uses random hyperplanes not necessarily orthogonal to the coordinate axes to partition the data while growing the isolation trees. Considering random hyperplanes has been mentioned before in Liu et al. (2010) but not furthered, while these authors focused on optimizing the splitting value so as to improve the dectection of clustered anomalies. But contrary to the original isolation forest algorithm, the random hyperplanes generated by the method in Hariri et al. (2021) are not guaranteed to actually produce a partition of the data, thus leading to empty branches in the isolation trees. This is observed and remedied in Lesouple et al. (2021) who propose a modification of the random hyperplane generation algorithm to ensure that data are present on both sides of the hyperplane. Among the other variants of the isolation forest method that have been introduced, Karczmarek et al. (2020) propose to use the k-means algorithm to build k-ary trees instead of binary trees. While most extensions have focused on modifying the splitting mechanism, Mensi and Bicego (2021) propose to weight the edges of the isolation trees, by accounting for the number of data points at each node, and they also introduce heuristics for aggregating the isolation trees in a forest, leading to a modified scoring function using those weights. Another scoring strategy which relies on a majority vote instead of on an average is proposed in Chabchoub et al. (2022). Mensi et al. (2023) introduce an adaptation of the isolation forest method to the case of pairwise data, and an extension to functional data is proposed in Staerman et al. (2019).

Despite its wide use and various extensions, the statistical properties of the isolation forest method remain little understood. Recall that the isolation forest anomaly detection method is a two-stage procedure, comprising a training stage, where a forest of isolation trees is constructed

from a training sample, followed by a scoring stage, where the induced scoring function is evaluated on the test data that may, or may not, differ from the training data. The regions of abnormality produced by an isolation forest then coincide with the upper level sets of the scoring function. One first question of interest is to determine, and analyze, the scoring function that is induced by the isolation forest method over a given finite sample at the population level, meaning at the limit when the number of trees tends to infinity. In this direction, we are only aware of the recent work of Morales et al. (2022), where they suggest an expression for the average heights analogous to the one that we obtain in our Theorem 5, as given by equation (11). Another central problem is to determine the behaviour of this limit scoring function in the large sample regime, that is to say when the number of data points tends to infinity. It is worth noticing that the large sample regime is not of mere theoretical interest but corresponds to the practical situation where the anomaly detector is trained with a large number of data that are known to be non-abnormal before being applied next to new, unseen data.

In the present paper, we address these questions in non asymptotic and asymptotic settings. We focus first on the isolation forest methodology in dimension one. Building upon work of Seidel and Aragon (1996) on randomized search trees and treaps (see also Aragon & Seidel, 1989), we start by introducing a sequential procedure of construction of random trees which have the structure of a treap and that we relate to the isolation trees (Proposition 2). This connection between isolation trees and the treaps that we introduce facilitates the analysis and using this, we derive an analytical expression, as a function of the sample points, for the limit of the scoring function over a finite sample when the number of trees tends to infinity (Theorem 5). Based on the limit expression obtained in Theorem 5, we deduce that the isolation forest algorithm is effective at detecting geometrically isolated points, although the efficiency of the dectection may be affected by an effect of scale that we reveal. Next, we study the large sample limit of the scoring function, as the number of data points tends to infinity. In this asymptotic regime, we consider a random design where points are drawn from an underlying distribution, as well as sequences of fixed designs. The results that we obtain (Theorem 7, Theorem 8, Corollary 9, Theorem 10 and Theorem 11) imply that asymptotically the isolation forest method operates as a detector of the support of the underlying distribution in the same way the method of one-class support vector machines does, although the design principles differ since one-class support vector machine are introduced for the specific goal of support recovery (Schölkopf et al., 1999, 2001). In fact, our initial intuition was that the scoring function would converge, after proper scaling, towards some kind of data depth function (see for instance Mosler, 2013) or at least to a function depending on the underlying density. Our results reveal that this is not the case since the only dependence on the distribution that remains at the limit is through its support, and this holds whether the support is simply connected or multiply connected. This set of results also allows us to derive robustness properties of the isolation forest methodology when the training set is contaminated by anomalous data. We take an asymptotic stance and we consider dense and sparse regimes. In the dense regime, anomalies are clustered and dense at the limit, meaning that they aggregate in a cluster of positive density. In this case, anomalies are not detected as such by an isolation forest, a phenomenon known as the masking effect. On the other hand, in a sparse regime where the proportion of anomalies tend to zero sufficiently fast, the isolation forest method is found to be robust to contamination during training in the sense that the support of the normal data is correctly recovered at the limit.

The rest of the paper is organized as follows. In Section 2, we describe the isolation forest method and we introduce some notation. The sequential procedure of construction of random trees is introduced in Section 3. The analytical expression for the limit of the scoring function over a finite sample is exposed in Section 4 and is studied in Section 5 in a non asymptotic setting. In Section 6, we present the asymptotic results in the large sample regime, under random designs and fixed designs. We end with Section 7 where we give some concluding remarks and we make a link with the Hilbert kernel density estimate introduced in Devroye and

Krzyżak (1999). Section 8 is devoted to the proofs and several technical results are gathered in an Appendix, at the end of the paper.

2 Isolation forest

In this section, we describe the isolation forest algorithm and we introduce some notation and vocabulary. Following Liu et al. (2008), an *isolation tree* over $\mathbb{R}^d$ is a binary tree that is grown using a given data set and that represents a binary recursive partition of $\mathbb{R}^d$. We start with the definition of a binary tree, which we take as a rooted, ordered and labelled tree in which each node has at most two children. Next we define a binary recursive partitition of a given set. Then we formalize the notion of an isolation tree.

Let $\mathcal{U} = \bigcup_{n\geq 0}\{0,1\}^n$ be the set of labels with the convention that $\{0,1\}^0 = \{\emptyset\}$. An element of $\mathcal{U}$ is a tuple of the form $u = (u^1, \dots, u^n)$ and its length is denoted by $|u| = n$ with the convention $|\emptyset| = 0$. Given $u = (u^1, \dots, u^m) \in \mathcal{U}$ and $v = (v^1, \dots, v^n) \in \mathcal{U}$, we write $uv = (u^1, \dots, u^m, v^1, \dots, v^n)$ for the concatenation of u and v, with the convention that $\emptyset u = u\emptyset = u$. The element $\emptyset$ is called the *root* and the elements of the form $u0$ and $u1$ are called the *left and right children* of u, respectively. A *binary tree* $\mathcal{T}$ is a finite subset of $\mathcal{U}$ such that:

(i) $\emptyset \in \mathcal{T}$,

(ii) $uv \in \mathcal{T} \implies u \in \mathcal{T}$.

A binary tree $\mathcal{T}$ is called *proper*, or *full*, if it satisfies the property $u0 \in \mathcal{T} \iff u1 \in \mathcal{T}$. Thus each node of a binary tree has at most two children while each node of a proper binary tree has either 0 or 2 children. A node without children is called a *leaf*, and the other nodes are called *internal nodes*. The set of leaves of a tree $\mathcal{T}$ is denoted by $\partial\mathcal{T}$. Given a tree $\mathcal{T} \neq \{\emptyset\}$, we denote by $\mathcal{T}^\circ = \mathcal{T} \setminus \partial\mathcal{T}$ the tree composed of the internal nodes of $\mathcal{T}$. For each $n \geq 1$, we denote by $\mathcal{B}_n$ the set of proper binary trees of size n, where the *size* of a tree is defined as its total number nodes, including the root. Given a tree $\mathcal{T}$, the *height* of a node $u \in \mathcal{T}$ is defined by its length, meaning that it is equal to $|u|$. From the perspective of graph theory (see for instance Diestel, 2017), a binary tree, as defined here, is a connected acyclic graph $(V(\mathcal{T}), E(\mathcal{T}))$ with a special vertex (the root), where any vertex has at most two children, and that is ordered and labelled (left and right children are distinguished even when a node has only one child). Due to this distinction, the definition of a binary tree that we use here is more convenient for our purposes, but we retain some concepts from graph theory. In particular, the height of a node $u \in \mathcal{T}$ corresponds to the length of the (unique) shortest path from the root of the tree to u, where the path length is defined as the number of edges in the path.

Given a set $\mathcal{S}$, we define a *binary recursive partition* of $\mathcal{S}$ as a pair $(\mathcal{T}, \pi_\mathcal{T})$, where $\mathcal{T}$ is a proper binary tree, and where $\pi_\mathcal{T} : \mathcal{T} \to \mathcal{P}(\mathcal{S})$ is a function such that $\pi_\mathcal{T}(\emptyset) = \mathcal{S}$, and such that $\{\pi_\mathcal{T}(v0), \pi_\mathcal{T}(v1)\}$ is a partition of $\pi_\mathcal{T}(v)$ for any internal node $v \in \mathcal{T}^\circ$, where $\mathcal{P}(\mathcal{S})$ denotes the power set of $\mathcal{S}$. The elements $\pi_\mathcal{T}(v)$, for $v \in \mathcal{T}$, will be called *cells* and we note that the collection $\{\pi_\mathcal{T}(v) : v \in \partial\mathcal{T}\}$ of cells associated with the leaves of $\mathcal{T}$ forms a partition of $\mathcal{S}$. We denote by $\mathfrak{P}(\mathcal{S})$ the set of all binary recursive partitions of $\mathcal{S}$. It will be convenient to restrict a partition to some subset. Given a subset $\mathcal{S}' \subset \mathcal{S}$ and a binary recursive partition $(\mathcal{T}, \pi_\mathcal{T})$ on $\mathcal{S}$, we define the *restriction of* $(\mathcal{T}, \pi_\mathcal{T})$ *to* $\mathcal{S}'$, as the recursive partition $(\mathcal{T}', \pi_{\mathcal{T}'})$ where $\mathcal{T}'$ is the subtree of $\mathcal{T}$ defined by

$$\mathcal{T}' = \{v \in \mathcal{T} : \pi_\mathcal{T}(v) \cap \mathcal{S}' \neq \emptyset\}, \tag{1}$$

and where $\pi_{\mathcal{T}'} : \mathcal{T}' \to \mathcal{P}(\mathcal{S})$ is defined by

$$\pi_{\mathcal{T}'}(v) = \pi_\mathcal{T}(v) \cap \mathcal{S}', \quad \text{for } v \in \mathcal{T}'. \tag{2}$$

By an *isolation tree over* $\mathbb{R}^d$ we mean a binary recursive partition $(\mathcal{T}, \pi_\mathcal{T})$ of $\mathbb{R}^d$ that is designed to isolate each data point at its leaves. Liu et al. (2008) define isolation trees with

cells obtained by partitioning along the coordinate axis. Then it holds $\pi(\emptyset) = \mathbb{R}^d$, and for any internal node $v \in \mathcal{T}^\circ$, the cells associated with the left and right children of v may be expressed as

$$\begin{cases} \pi_{\mathcal{T}}(v0) = \pi_{\mathcal{T}}(v) \cap \left\{x \in \mathbb{R}^d \,:\, x^{(j)} \leq \tau\right\}, \\ \pi_{\mathcal{T}}(v1) = \pi_{\mathcal{T}}(v) \cap \left\{x \in \mathbb{R}^d \,:\, x^{(j)} > \tau\right\}, \end{cases} \tag{3}$$

for some pair (j, τ) composed of a *component number* $j \in \{1, \ldots, d\}$ and of a *split value* $\tau \in \mathbb{R}$, and where $x^{(j)}$ denotes the j^{th} component of $x \in \mathbb{R}^d$. When no risk of confusion may arise, we may simply denote an isolation tree $(\mathcal{T}, \pi_{\mathcal{T}})$ as $\mathcal{T}$ for ease of notation.

Let $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ be a data set composed of n points in $\mathbb{R}^d$. The isolation trees that we consider are grown recursively according to the following procedure. The structure is initialized with $\mathcal{T}$ composed only of the root and with associated cell $\mathbb{R}^d$. Next, a pair (j, τ) is first generated from $\mathcal{D}_n$ by a random draw of a component number j uniformly among $\{1, \ldots, d\}$, followed by a random draw of a split value τ uniformly over the interval $[\min\{x_i^{(j)} \,:\, 1 \leq i \leq n\}, \max\{x_i^{(j)} \,:\, 1 \leq i \leq n\}]$. The two children $\emptyset 0$ and $\emptyset 1$ of the root are then inserted in $\mathcal{T}$ and $\pi_{\mathcal{T}}(\emptyset 0)$ and $\pi_{\mathcal{T}}(\emptyset 1)$ are defined according to (3). Next, $\mathcal{D}_n$ is partitioned into $\mathcal{D}_n^{(0)} = \mathcal{D}_n \cap \pi_{\mathcal{T}}(\emptyset 0)$ and $\mathcal{D}_n^{(1)} = \mathcal{D}_n \cap \pi_{\mathcal{T}}(\emptyset 1)$ and the left and right subtrees of the root node are grown recursively in a similar manner using $\mathcal{D}_n^{(0)}$ and $\mathcal{D}_n^{(1)}$ respectively. The recursion on a subtree ends either when the data set resulting from the sequence of splits contains only one data point, or when a height limit is reached. When an isolation tree is fully grown, each cell associated with the leaves of the tree contains exactly one data point; thus, in this case, the data points are isolated by the tree, which then has n leaves, $n-1$ internal nodes, and so a size of $2n-1$. The procedure is summarized in Algorithm 1. We also note that, if the projections of the data points along the coordinate axes are all distinct, then there are as many different isolation tree structures as there are (proper) binary trees in $\mathcal{B}_{2n-1}$, meaning that the application $(\mathcal{T}, \pi_{\mathcal{T}}) \mapsto \mathcal{T}$ from $\mathbb{T}$ to $\mathcal{B}_{2n-1}$ is surjective, and the cardinality of $\mathcal{B}_{2n-1}$ is known to be $\frac{1}{n}\binom{2(n-1)}{n-1}$, as shown for instance in Drmota (2009, Theorem 2.1).

Given an isolation tree $\mathcal{T}$ over $\mathbb{R}^d$, let $h_{\mathcal{T}}$ be the piecewise-constant function defined for any $x \in \mathbb{R}^d$ by

$$h_{\mathcal{T}}(x) = \sum_{v \in \partial\mathcal{T}} |v| \mathbf{1}_{\pi_{\mathcal{T}}(v)}(x). \tag{4}$$

Let $\mathcal{T}_1, \ldots, \mathcal{T}_N$ be an ensemble of N isolation trees constructed independently from the data set $\mathcal{D}_n$ according to algorithm 1. Liu et al. (2008) define an anomaly score $s_N(x)$ at x by

$$s_N(x) = 2^{-\frac{1}{c(n)N}\sum_{\ell=1}^{N} h_{\mathcal{T}_\ell}(x)}, \tag{5}$$

where $c(n)$ is the average path length of unsuccessful searches in a binary search tree and is taken as $c(n) = 2\mathcal{H}_{n-1} - 2(n-1)/n$, where $\mathcal{H}_\ell = \sum_{k=1}^{\ell} \frac{1}{k}$ denotes the ℓ^{th} harmonic number, for $\ell \geq 1$. Liu et al. (2008) also propose growing the isolation trees from random subsamples of $\mathcal{D}_n$ of size $m \leq n$, in which case $c(n)$ is replaced by $c(m)$ in (5). The rationale behind the isolation forest algorithm is that anomalies are more suceptible of being isolated earlier in the recursive partitioning procedure than non abnormal data points. Thus anomalies are expected to be isolated at shorter heights on average, thereby receiving a score close to 1, while non abnormal data points are expected to receive lower score values.

Remark 1. *In the definitions given above, two properties are implicitely assumed to hold. The first one is that the data points in $\mathcal{D}_n$ may all be isolated by an isolation tree grown to maximal height. The second one is that the pair (j, τ) in step 3 of Algorithm 1 produces a proper split, meaning that it actually creates two children, and this is guaranteed with probability one when $\min\{x_i^{(j)} \,:\, x_i \in \mathcal{S}\} < \max\{x_i^{(j)} \,:\, x_i \in \mathcal{S}\}$. These two properties are satisfied, for instance, under the condition that the projections of the data points in $\mathcal{D}_n$ over each coordinate axis are distinct, and this holds with probability one when the data is drawn from a continuous distribution over*

$\mathbb{R}^d$. *When the data are only assumed to be distinct (but the projections on some coordinate axis may not be distinct), then Algorithm 1 may be applied by modifying step 2 so that j is chosen among those components for which the projections $\{x_i^{(j)} : x_i \in \mathcal{S}\}$ are not all equal (note that the two conditions are equivalent in dimension 1). We use this variant in Proposition 6 and in our analysis in Section 6.2. As pointed out by a referee, this setting may result from a discretization of the data, even leading to duplicated values in $\mathcal{D}_n$. In the case of duplicated data, Algorithm 1 may be applied with the additional condition that the recursion on a leaf node stops when either that node contains only one data point, or when all the data at that node are identical, which is in line with the algorithm in Liu et al. (2008), and this results in the same distribution of scoring functions as the one that is produced by applying Algorithm 1 without modification to the subset of distinct data from $\mathcal{D}_n$. Thus it is sufficient to assume that all the data in $\mathcal{D}_n$ are distinct, and the analysis that we develop also applies to the setting of duplicated values with the effect of reducing the sample size from n to the number of distinct values in $\mathcal{D}_n$.*

Algorithm 1 Recursive definition of an isolation tree $\mathcal{T}$ on a finite sample of size n. Nodes of $\mathcal{T}$ are denoted by v and their associated cell by $\pi_{\mathcal{T}}(v)$. The tree is grown until all n points are isolated (this holds under minimal assumptions on $\mathcal{D}_n$; see Remark 1).

Input: Point set $\mathcal{D}_n = \{x_1, \dots, x_n\}$ in $\mathbb{R}^d$.
Output: An isolation tree $(\mathcal{T}, \pi_{\mathcal{T}})$.

Initialization: Set $\mathcal{T} = \{\emptyset\}$ and $\pi_{\mathcal{T}}(\emptyset) = \mathbb{R}^d$.

Recursion on a leaf node v of $\mathcal{T}$:
1: Let $S = \pi_{\mathcal{T}}(v) \cap \mathcal{D}_n$.
if $\mathcal{S}$ contains more than one point **then**
  2: Draw a component $j \in \{1, \dots, d\}$ uniformly.
  3: Draw a split point τ uniformly in the interval $\left[\min\{x_i^{(j)} : x_i \in \mathcal{S}\}, \max\{x_i^{(j)} : x_i \in \mathcal{S}\}\right]$.
  4: Insert nodes $v0$ and $v1$ in $\mathcal{T}$ as left and right children of v respectively.
  5: Set $\pi_{\mathcal{T}}(v0) = \pi_{\mathcal{T}}(v) \cap \left\{x \in \mathbb{R}^d : x^{(j)} \leq \tau\right\}$ and $\pi_{\mathcal{T}}(v1) = \pi_{\mathcal{T}}(v) \cap \left\{x \in \mathbb{R}^d : x^{(j)} > \tau\right\}$.
  6: Apply this recursion to $v0$.
  7: Apply this recursion to $v1$.
end if

Let the data set $\mathcal{D}_n$ be fixed. We denote by $\mathbb{T} \subset \mathfrak{P}(\mathbb{R}^d)$ the set of all possible isolation trees that may be grown from $\mathcal{D}_n$ using Algorithm 1 (the dependence on $\mathcal{D}_n$ is understood) and by μ be the probability measure that is induced by Algorithm 1 over $\mathbb{T}$. When the objective is to detect anomalies within $\mathcal{D}_n$, the scoring function needs to be evaluated at the data points only. It will be convenient to introduce isolation trees restricted to $\mathcal{D}_n$ for this purpose, as well as for the analysis of the distributional properties of the isolation forest methodology in general, where the restriction to a subset is defined in (1) and (2). Indeed, we note that each isolation tree yields a binary recursive partition of $\mathcal{D}_n$ which is obtained by filtering down $\mathcal{D}_n$ through the tree, as illustrated in Figure 1.

Let $\Pi_n : \mathbb{T} \to \mathfrak{P}(\mathcal{D}_n)$ be the application mapping each isolation tree $(\mathcal{T}, \pi_{\mathcal{T}}) \in \mathbb{T}$ to its restriction to $\mathcal{D}_n$, and let $\mathbb{T}_n = \{\Pi_n((\mathcal{T}, \pi_{\mathcal{T}})) : (\mathcal{T}, \pi_{\mathcal{T}}) \in \mathbb{T}\}$ be the set of all such restricted isolation trees. Notice that $(\mathcal{T}, \pi_{\mathcal{T}})$ and $\Pi_n((\mathcal{T}, \pi_{\mathcal{T}}))$ carry the same tree structure, so that $\Pi_n((\mathcal{T}, \pi_{\mathcal{T}})) = (\mathcal{T}, \pi_{\mathcal{T},n})$, and where $\pi_{\mathcal{T},n} : \mathcal{T} \to \mathcal{P}(\mathcal{D}_n)$ is defined according to (2). Indeed, this is due to the fact that $\pi_{\mathcal{T}}(v) \cap \mathcal{D}_n \neq \emptyset$ for all $v \in \mathcal{T}$ since the recursive growth of $\mathcal{T}$ by Algorithm 1 is stopped before the cells associated with the leaves are empty of data points. Given $(\mathcal{T}, \pi_{\mathcal{T},n}) \in \mathbb{T}_n$, we denote by $h_{\mathcal{T},n} : \mathcal{D}_n \to \mathbb{N}$ the height function defined as in (4),

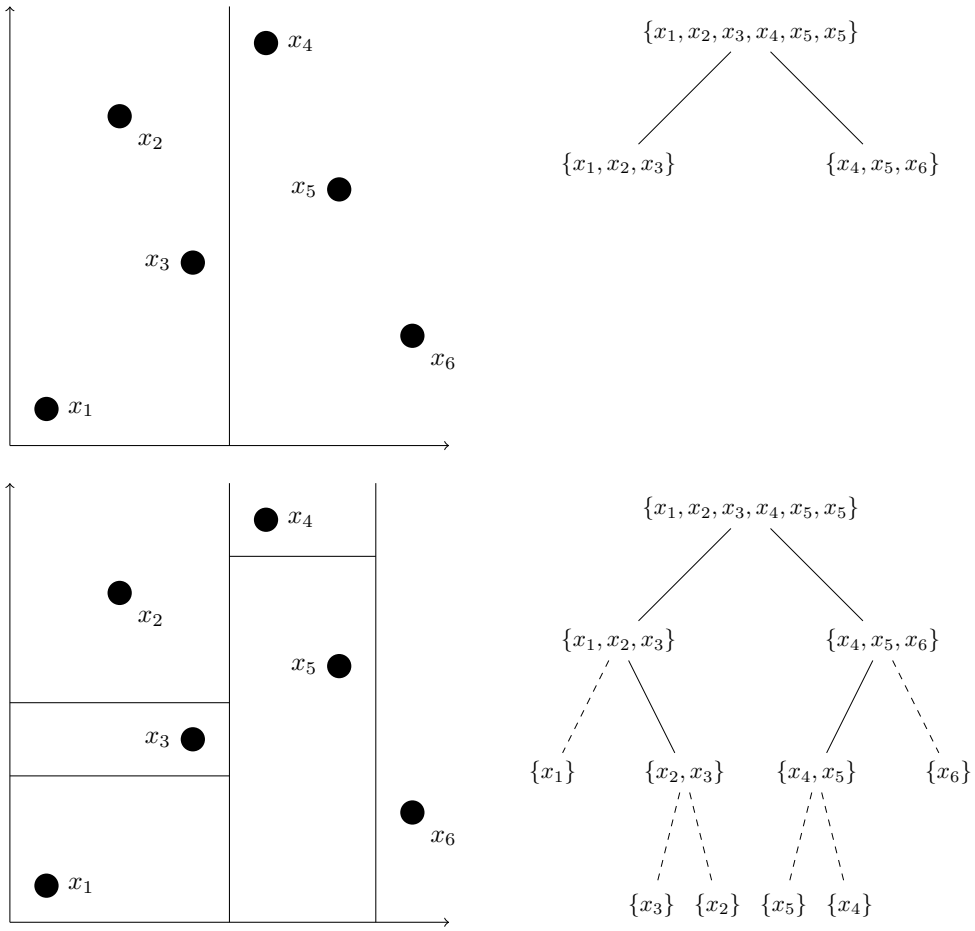


Figure 1: Construction of an isolation tree on 6 data points in $\mathbb{R}^2$. Top row: induced partition (left) and isolation tree (right) after the first split. Bottom row: final partition and corresponding isolation tree.

meaning that

$$h_{\mathcal{T},n}(x) = \sum_{v \in \partial \mathcal{T}} |v| \mathbf{1}_{\pi_{\mathcal{T},n}}(x), \quad \text{for } x \in \mathcal{D}_n, \tag{6}$$

and we note that

$$(h_{\mathcal{T}}(x_1), \dots, h_{\mathcal{T}}(x_n)) = \left(h_{\mathcal{T},n}(x_1), \dots, h_{\mathcal{T},n}(x_n)\right),$$

for any $(\mathcal{T}, \pi_{\mathcal{T}}) \in \mathbb{T}$ such that $(\mathcal{T}, \pi_{\mathcal{T},n}) = \Pi_n\left((\mathcal{T}, \pi_{\mathcal{T}})\right)$. We denote by $\mu_n = \mu \circ \Pi_n^{-1}$ the image measure of μ by Π_n on $\mathbb{T}_n$.

3 Isolation tree: from recursive to sequential construction

In this section, we focus on isolation trees in dimension 1. The setting is that of a data set $\mathcal{D}_n = \{x_1, \dots, x_n\}$ composed of n distinct points in $\mathbb{R}$ indexed in increasing order $x_1 < x_2 < \cdots < x_n$. For $i = 1, \dots, n-1$, we let $I_i = [x_i, x_{i+1})$ save for $I_{n-1} = [x_{n-1}, x_n]$, and we denote by $w_i = x_{i+1} - x_i$ its length. We consider isolation trees grown using Algorithm 1 and that isolate each data point of $\mathcal{D}_n$ at their leaves.

In dimension 1, any isolation tree restricted to $\mathcal{D}_n$ satisfies the property that its partition map is completely determined by its tree. Indeed, consider the generation of an isolation tree according to Algorithm 1 and suppose that the first split point (used to partition $\mathcal{D}_n$) falls in

the interval I_k, for some $k \in \{1, \dots, n-1\}$. Then the cell of the left child of the root contains the first k points of $\mathcal{D}_n$ and the cell of the right child of the root contains the remaining $n-k$ points. Conversely, for any $\mathcal{D}_n$-restricted isolation tree $(\mathcal{T}, \pi_{\mathcal{T},n}) \in \mathbb{T}_n$, if the left subtree of the root of $\mathcal{T}$ contains k leaves, for some $k \in \{1, \dots, n-1\}$, then necessarily $\pi_{\mathcal{T},n}(\emptyset 0) = \{x_1, \dots, x_k\}$ and $\pi_{\mathcal{T},n}(\emptyset 1) = \{x_{k+1}, \dots, x_n\}$. By recursion on the left and right subtrees, this shows that $\pi_{\mathcal{T},n}$ is completely determined by $\mathcal{T}$. Consequently, in dimension 1, $\mathbb{T}_n$ can be identified with $\mathcal{B}_{2n-1}$ through the application $\iota : \mathbb{T}_n \to \mathcal{B}_{2n-1}$ defined by $\iota\left((\mathcal{T}, \pi_{\mathcal{T},n})\right) = \mathcal{T}$ which is bijective in dimension 1 (but ι fails to be injective in dimension larger than 1 due to the fact that $\mathcal{T}$ does not carry information about the coordinate axis along which the cells are partitioned). Using this observation together with ideas introduced in Seidel and Aragon (1996) for the study of randomized search trees, we define a sequential procedure that generates randomized binary search trees with set of keys $\{1, \dots, n-1\}$ and we prove that the two procedures (recursive and sequential) generate the same distribution of trees in a sense made precise in Proposition 2.

A binary search tree with set of keys $\{1, \dots, n-1\}$ is a binary tree $\mathcal{T} \in \mathcal{B}_{2n-1}$ together with a bijective application $L_{\mathcal{T}} : \mathcal{T}^\circ \to \{1, \dots, n-1\}$ that gives the keys stored at the internal nodes of $\mathcal{T}$, and that satisfies the binary search tree property that, for any $u \in \mathcal{T}$, $L_{\mathcal{T}}(v) < L_{\mathcal{T}}(u)$ for any $v \in \mathcal{T}^\circ$ that belongs to the left subtree of u, and $L_{\mathcal{T}}(v) > L_{\mathcal{T}}(u)$ for any $v \in \mathcal{T}^\circ$ that belongs to the right subtree of u. We denote by $\mathcal{B}^S_{n-1}$ the set of all binary search trees with set of keys $\{1, \dots, n-1\}$. In fact, when the set of keys is $\{1, \dots, n-1\}$, as we consider here, $L_{\mathcal{T}}$ is completely determined by $\mathcal{T}$, so that each binary search tree $(\mathcal{T}, L_{\mathcal{T}})$ is canonically identified with the element $\mathcal{T}$ of $\mathcal{B}_{2n-1}$, as we argue below. Binary search trees $(\mathcal{T}(\alpha), L_{\mathcal{T}(\alpha)})$ are generated by a permutation α of $\{1, \dots, n-1\}$. We use the recursive description given in Drmota (2009, chapter 1). Let $\mathcal{K} = \{1, \dots, n-1\}$ and let $k_0 = \min \mathcal{K}$ (so that $k_0 = 1$). The construction of $\mathcal{T}(\alpha)$ starts with the root $\emptyset$ and the assignment $L_{\mathcal{T}(\alpha)}(\emptyset) = \alpha(k_0)$. Next, $\alpha(k_0)$ is taken as a pivot to partition $\{2, \dots, n-1\}$ into $\mathcal{K}_0 = \{2 \le k \le n-1 \,:\, \alpha(k) < \alpha(k_0)\}$ and $\mathcal{K}_1 = \{2 \le k \le n-1 \,:\, \alpha(k) > \alpha(k_0)\}$. This procedure is applied recursively to build the left and right subtrees based on $\mathcal{K}_0$ and on $\mathcal{K}_1$ respectively. Once the recursion completes, the tree reaches the size of $n-1$ and finally, n leaves are added wherever possible to $\mathcal{T}(\alpha)$ so as to make $\mathcal{T}(\alpha)$ a proper binary tree of size $2n-1$. Hence the procedure generates $\mathcal{T}(\alpha) \in \mathcal{B}_{2n-1}$ together with the bijective application $L_{\mathcal{T}(\alpha)} : \mathcal{T}(\alpha)^\circ \to \{1, \dots, n-1\}$. Notice that the generation by permutation produces as many binary search trees as there are elements in $\mathcal{B}^S_{n-1}$, meaning that the application $\alpha \mapsto \mathcal{T}_\alpha$ from the set of permutations of $\{1, \dots, n-1\}$ to $\mathcal{B}^S_{n-1}$ is surjective. It is also worth noticing that the application $\tilde{\iota} : \mathcal{B}^S_{n-1} \to \mathcal{B}_{2n-1}$ defined by $\tilde{\iota}\left((\mathcal{T}, L_{\mathcal{T}})\right) = \mathcal{T}$ is bijective. To be sure, for each $\mathcal{T} \in \mathcal{B}_{2n-1}$, denote by $k_0(v)$ and $k_1(v)$ the number of internal nodes (including v) in the left and right subtrees at $v \in \mathcal{T}^\circ$ respectively, and let $L^{\mathcal{B}}_{\mathcal{T}} : \mathcal{T}^\circ \to \{1, \dots, n-1\}$ be the labelling function defined recursively by

$$\begin{cases} L^{\mathcal{B}}_{\mathcal{T}}(v0) = L^{\mathcal{B}}_{\mathcal{T}}(v) - k_0(v) + k_0(v0), & \text{for } v0 \in \mathcal{T}^\circ, \\ L^{\mathcal{B}}_{\mathcal{T}}(v1) = L^{\mathcal{B}}_{\mathcal{T}}(v) + k_1(v) - k_1(v1), & \text{for } v1 \in \mathcal{T}^\circ, \end{cases} \tag{7}$$

with the initial condition that $\mathcal{L}^{\mathcal{B}}_{\mathcal{T}}(\emptyset) = k_0(\emptyset)$. Then we see that $L^{\mathcal{B}}_{\mathcal{T}}$ is bijective and satisfies the binary search tree property, implying that $(\mathcal{T}, L^{\mathcal{B}}_{\mathcal{T}}) \in \mathcal{B}^S_{n-1}$ and so that $\tilde{\iota}$ is surjective. Next, for any $(\mathcal{T}, L_{\mathcal{T}})$ and $(\mathcal{T}', L_{\mathcal{T}'})$ in $\mathcal{B}^S_{n-1}$ such that $\tilde{\iota}\left((\mathcal{T}, L_{\mathcal{T}})\right) = \tilde{\iota}\left((\mathcal{T}', L_{\mathcal{T}'})\right)$, we then have $\mathcal{T} = \mathcal{T}'$ and since $L_{\mathcal{T}}$ and $L_{\mathcal{T}'}$ both satisfy the binary search tree property, we necessarily have $L_{\mathcal{T}}(\emptyset\eta) = L^{\mathcal{B}}_{\mathcal{T}}(\emptyset\eta) = L_{\mathcal{T}'}(\emptyset\eta)$, for any $\eta \in \{0,1\}$, implying by recursion on the left and right subtrees that $L_{\mathcal{T}} = L_{\mathcal{T}'}$, and so that $\tilde{\iota}$ is injective.

We now introduce the following randomization of binary search trees with keys $\{1, \dots, n-1\}$. Let F_0 be the cumulative distribution function of the uniform distribution over $[0,1]$ (the particular choice of distribution does not really matter as long as it is absolutely continuous). Let $Z_1, \dots, Z_{n-1}$ be independent random variables where Z_i is distributed according to $F_0^{w_i}$, for $i = 1, \dots, n-1$. Assuming no ties among the Z_i's (since this occurs with probability one), let α be the permutation of $\{1, \dots, n-1\}$ such that $Z_{\alpha(1)} \ge Z_{\alpha(2)} \ge \dots Z_{\alpha(n-1)}$. Then we consider

Algorithm 2 Sequential generation of the random binary search trees with set of keys $\{1, \dots, n-1\}$.

Input: Random variables $Z_1, \dots, Z_{n-1}$.
Output: Randomized binary search tree $(\mathcal{T}, L_{\mathcal{T}})$ with set of keys $\{1, \dots, n-1\}$.

1: Sort the Z_i's by decreasing order $Z_{\alpha(1)} > Z_{\alpha(2)} > \cdots > Z_{\alpha(n-1)}$.
2: Set $\mathcal{T}' = \{\emptyset\}$ and $L_{\mathcal{T}'}(\emptyset) = \alpha(1)$.
3: Sequential insertion of the internal nodes:
for $j = 2, \dots, n-1$ **do**
  Set the current node v to the root $\emptyset$.
  while $[(v0 \in \mathcal{T}$ and $\alpha(j) < L_{\mathcal{T}}(v))$ or $(v1 \in \mathcal{T}$ and $\alpha(j) > L_{\mathcal{T}}(v))]$ **do**
    if $\alpha(j) < L_{\mathcal{T}}(v)$ **then**
      $v \leftarrow v0$
    else
      $v \leftarrow v1$
    end if
  end while
  if $\alpha(j) < L_{\mathcal{T}}(v)$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{v0\}$ and set $L_{\mathcal{T}}(v0) = \alpha(j)$.
  else
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{v1\}$ and set $L_{\mathcal{T}}(v1) = \alpha(j)$
  end if
end for
4: Add n leaves to complete the tree: $\mathcal{T} \leftarrow \mathcal{T} \cup \bigcup_{v \in \mathcal{T}} (\{v0\} \cup \{v1\})$.

the randomized binary search tree $(\mathcal{T}, L_{\mathcal{T}})(Z_1, \dots, Z_{n-1}) := \left(\mathcal{T}(\alpha), L_{\mathcal{T}(\alpha)}\right)$ generated from the random permutation α of $\{1, \dots, n-1\}$. The procedure is summarized in Algorithm 2, where we use the equivalent formulation of generation of a binary search tree by sequential insertion of the nodes (see Drmota, 2009, chapter 1). This procedure defines a probability measure on $\mathcal{B}^S_{n-1}$ that we denote by $\tilde{\mu}_n$. Notice that when the interval lengths $\{w_\ell : 1 \leq \ell \leq n-1\}$ are all equal, then all permutations are equally likely so that $\tilde{\mu}_n$ corresponds to the distribution of the standard probabilistic model of binary search trees that are generated by uniform permutations of the keys.

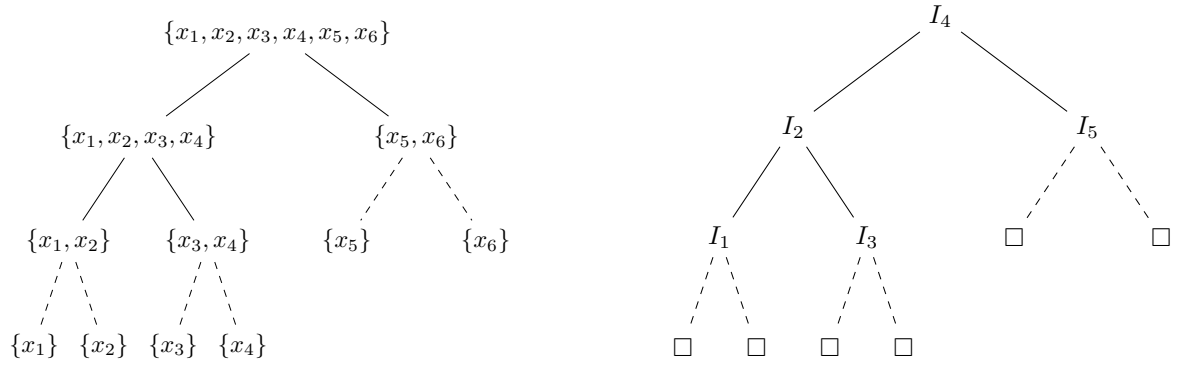
We may now formalize the connection between the two probability measures induced by the recursive and sequential random generation procedures.

Proposition 2. *The image measure of μ_n by $\tilde{\iota}^{-1} \circ \iota$ is equal to $\tilde{\mu}_n$, that is to say, it holds $\tilde{\mu}_n = \mu_n \circ (\iota^{-1} \circ \tilde{\iota})$.*

Hence since $\tilde{\iota}^{-1} \circ \iota : \mathbb{T}_n \to \mathcal{B}^S_{n-1}$ is bijective, the properties of the distibution of the isolation trees restricted to $\mathcal{D}_n$ may be deduced from those of the binary search trees. Given a $\mathcal{D}_n$-restricted isolation tree $(\mathcal{T}, \pi_{\mathcal{T},n}) \in \mathbb{T}_n$, $(\tilde{\iota}^{-1} \circ \iota)(\mathcal{T}, \pi_{\mathcal{T},n})$ is the binary search tree in $\mathcal{B}^S_{n-1}$ which stores in its nodes the indices of the intervals that contain the split points that generate the recursive partition, as illustrated in Figure 2.

4 The isolation forest average heights

Equipped with Proposition 2, we study the properties of the height function of isolation trees by first relating it with the height function of the nodes of the binary search trees. Next in Theorem 5, we establish the analytical expressions for the expectation of the heights of the leaves of a random $\mathcal{D}_n$-restricted isolation tree $(\mathcal{T}, \pi_{\mathcal{T},n})$ distributed as μ_n. The setting is the



(a) Isolation tree $\mathcal{T}$

(b) Binary search tree $\mathcal{T}'$

Figure 2: An isolation tree $\mathcal{T}$ is represented in (a) and its associated binary search tree $\mathcal{T}' = (\tilde{\iota}^{-1} \circ \iota)(\mathcal{T})$ is represented in (b). Edges to the leaves are represented with dashed lines and the leaves of the binary search tree $\mathcal{T}'$ are represented with empty boxes.

one of the previous section, meaning that we consider a finite sample $x_1 < \cdots < x_n$. Let us emphasize that the data points are ordered and fixed (deterministic). In the case of a random sample, the expressions in Theorem 5 remain valid when expectations are replaced by conditional expectations on the sample. This is used in the asymptotic analysis that we develop in Section 6.

For any $\mathcal{T}^{IT} := (\mathcal{T}, \pi_{\mathcal{T},n}) \in \mathbb{T}_n$, denote by $H(\mathcal{T}^{IT}) = \big(h_{\mathcal{T},n}(x_1), \dots, h_{\mathcal{T},n}(x_n)\big)$ the vector the components of which are the heights of the n leaves of $\mathcal{T}$, where the height function $h_{\mathcal{T},n}$ is defined in (6). Given a binary search tree $\mathcal{T}^S := (\widetilde{\mathcal{T}}, L_{\widetilde{\mathcal{T}}}) \in \mathcal{B}^S_{n-1}$, let $D(\mathcal{T}^S) = (D_1(\mathcal{T}^S), \dots, D_{n-1}(\mathcal{T}^S))$, where $D_i(\mathcal{T}^S)$ is the height of the internal node of $\widetilde{\mathcal{T}}$ where the key i is stored, meaning that $D_i(\mathcal{T}^S) = \sum_{v \in \widetilde{\mathcal{T}}^\circ} |v| \mathbf{1}\{L_{\widetilde{\mathcal{T}}}(v) = i\}$. From the definitions of the bijections $\iota : \mathbb{T}_n \to \mathcal{B}_{2n-1}$ and $\tilde{\iota} : \mathcal{B}^S_{n-1} \to \mathcal{B}_{2n-1}$, it follows that whenever $\mathcal{T}^S = (\tilde{\iota}^{-1} \circ \iota)(\mathcal{T}^{IT})$, the two height functions are related by

$$h_{\mathcal{T},n}(x_i) = \begin{cases} 1 + D_1(\mathcal{T}^S) & \text{if } i = 1, \\ 1 + \max\big\{D_{i-1}(\mathcal{T}^S), D_i(\mathcal{T}^S)\big\} & \text{if } 2 \le i \le n-1, \\ 1 + D_{n-1}(\mathcal{T}^S) & \text{if } i = n, \end{cases} \tag{8}$$

which we write in condensed form as

$$H(\mathcal{T}^{IT}) = \Psi\left(D\left((\tilde{\iota}^{-1} \circ \iota)(\mathcal{T}^{IT})\right)\right), \tag{9}$$

for some adequate function $\Psi : \mathbb{R}^{n-1} \to \mathbb{R}^n$, and this holds for any $\mathcal{T}^{IT}$ in $\mathbb{T}_n$. Indeed, that (8) holds for boundary points (when $i = 1$ or $i = n$) is clear, and when $2 \le i \le n-1$, the point x_i becomes isolated only once split points have been drawn in both I_{i-1} and I_i. Implicitely, it is assumed here that no split point coincides exactly with one of the data points, which is not restrictive since this holds with probability one. Given a random $\mathcal{D}_n$-restricted isolation tree $\mathcal{T}^{IT}$ distributed as μ_n and a random binary search tree $\widetilde{\mathcal{T}}$ distributed as $\tilde{\mu}_n$, using (9) and applying Proposition 2 implies that $H(\mathcal{T}^{IT})$ has the same distribution as $\Psi(D(\mathcal{T}^S))$, thereby providing the link between the two random generation procedures.

Now we focus on the height function of the internal nodes of a binary search tree $\mathcal{T}^S := (\widetilde{\mathcal{T}}, L_{\widetilde{\mathcal{T}}}) \in \mathcal{B}^S_{n-1}$. We refer to each internal node $v \in \widetilde{\mathcal{T}}^\circ$ by the interval I_i for which $i = L_{\widetilde{\mathcal{T}}}(v)$. Following Seidel and Aragon (1996), the study of the heights D_i's is facilitated by the introduction of the binary *ancestor variables* $A_{ji} := A_{ji}(\mathcal{T}^S)$ defined by $A_{ji} = 1$ if node I_j is an ancestor of node I_i in $\widetilde{\mathcal{T}}$ and $A_{ji} = 0$ otherwise (the dependence of D_i and A_{ji} on $\widetilde{\mathcal{T}}$ is dropped

from the notation for clarity). We recall that a node I_j is said to be an ancestor of node I_i in $\widetilde{\mathcal{T}}$ if I_j belongs to the unique path from the root of $\widetilde{\mathcal{T}}$ to I_i and has a lower height than that of I_i. At this point, it is worth noting that for any $2 \leq i \leq n-1$, we always have $A_{i-1,i} + A_{i,i-1} = 1$, meaning that either node I_i is an ancestor of node I_{i-1} or node I_{i-1} is an ancestor of node I_i in $\widetilde{\mathcal{T}}$. Obviously they cannot be ancestors one of each other simultaneously, but that none of the two is an ancestor of the other cannot occur too. Indeed, if $A_{i-1,i} = A_{i,i-1} = 0$, then I_{i-1} and I_i belong to different subtrees. Denote by I_k their closest common ancestor, meaning the node with the largest height belonging to the intersection of the two unique paths from the root to I_{i-1} and I_i. Then $i-1 < k < i$ necessarily, hence a contradiction and so the relation $A_{i-1,i} + A_{i,i-1} = 1$ is true for any $2 \leq i \leq n-1$. From this, it follows that the maximum term in (8) may be expressed as $\max\{D_{i-1}, D_i\} = D_{i-1}A_{i,i-1} + D_i A_{i-1,i}$.

Recall that the height D_i of I_i in $\widetilde{\mathcal{T}}$ is equal to the length of the (unique) path from the root of $\widetilde{\mathcal{T}}$ to I_i, and where the length is defined as the number of edges in this path. Hence D_i may be expressed in terms of the ancestor variables as $D_i = \sum_{j=1}^{n} A_{ji}$ and this leads to

$$\max\{D_{i-1}, D_i\} = \sum_{j=1}^{n} A_{j,i-1}A_{i,i-1} + \sum_{j=1}^{n} A_{ji}A_{i-1,i}. \tag{10}$$

Given a random binary search tree $\mathcal{T}^S := \mathcal{T}^S(Z_1, \dots, Z_{n-1})$ as defined in Algorithm 2, the following Lemma characterizes the ancestor variables in terms of the random variables $Z_1, \dots, Z_{n-1}$. It is proved in Seidel and Aragon (1996, Lemma 4.3) in the context of randomized search trees where it is called the ancestor lemma. The setting considered in Seidel and Aragon (1996) corresponds to ours when the interval lengths $\{w_\ell : 1 \leq \ell \leq n-1\}$, are integers. Here, we provide a statement tailored to our context and a proof in section 8 for completeness.

Lemma 3. *In any randomized binary search tree $\mathcal{T}^S := \mathcal{T}^S(Z_1, \dots, Z_{n-1})$ generated according to Algorithm 2, node I_i is an ancestor of node I_j if and only if Z_i is the largest among all the Z_k's for k comprised between i and j, included. Thus*

$$A_{ij} = \mathbf{1}\left\{Z_i \geq \max\{Z_k \;\; i \wedge j \leq k \leq i \vee j\}\right\} = \prod_{k=i\wedge j}^{i\vee j} \mathbf{1}\{Z_i \geq Z_k\}.$$

Using Lemma 3, we may further expand the expression for the maximum term $\max\{D_{i-1}, D_i\}$ in (10).

Lemma 4. *In the setting of Lemma 3, for any $2 \leq i \leq n-1$, we have*

$$\max\{D_{i-1}, D_i\} = 2 + \sum_{j \in \mathcal{J}_1} A_{j,i-1} + \sum_{j \in \mathcal{J}_2} A_{ji},$$

where $\mathcal{J}_1 = \{j : 1 \leq j \leq i-2\}$ and $\mathcal{J}_2 = \{j : i+1 \leq j \leq n-1\}$, and where we use the convention that a sum over an empty set is equal to 0.

Notice that the two sums in Lemma 4 are independent as a consequence of Lemma 3. Using the results above, we deduce the distribution and the analytical expressions for the expectations and the variances of the heights of a random isolation tree.

Theorem 5. *Let $\mathcal{T}^{IT} := (\mathcal{T}, \pi_{\mathcal{T}})$ be a random isolation tree distributed according to μ. Then for any $1 \leq i \leq n$, $h_{\mathcal{T}}(x_i)$ is distributed as*

$$h_{\mathcal{T}}(x_i) \overset{\mathcal{L}}{=} M_i + N_i,$$

where M_i and N_i are independent random variables such that $M_i = 0$ if $i = 1$ and M_i follows a Poisson Binomial distribution with probabilities of successes given by $\left\{\frac{x_{j+1} - x_j}{x_i - x_j} : 1 \leq j \leq i-1\right\}$

if $2 \leq i \leq n$, *and that* $N_i = 0$ *if* $i = n$ *and* N_i *follows a Poisson Binomial distribution with probabilities of successes given by* $\left\{ \frac{x_j - x_{j-1}}{x_j - x_i} \,:\, i+1 \leq j \leq n \right\}$ *if* $1 \leq i \leq n-1$. *In particular, we have*

$$\mathbb{E}[h_{\mathcal{T}}(x_i)] = \begin{cases} \displaystyle\sum_{j=2}^{n} \frac{x_j - x_{j-1}}{x_j - x_1} & \text{if } i = 1, \\ \displaystyle\sum_{j=1}^{i-1} \frac{x_{j+1} - x_j}{x_i - x_j} + \sum_{j=i+1}^{n} \frac{x_j - x_{j-1}}{x_j - x_i} & \text{if } 2 \leq i \leq n-1, \\ \displaystyle\sum_{j=1}^{n-1} \frac{x_{j+1} - x_j}{x_n - x_j} & \text{if } i = n, \end{cases} \tag{11}$$

and

$$\mathbb{V}[h_{\mathcal{T}}(x_i)] = \begin{cases} \displaystyle\sum_{j=2}^{n} \frac{x_j - x_{j-1}}{x_j - x_1} \left(1 - \frac{x_j - x_{j-1}}{x_j - x_1}\right) & \text{if } i = 1, \\ \displaystyle\sum_{j=1}^{i-1} \frac{x_{j+1} - x_j}{x_i - x_j} \left(1 - \frac{x_{j+1} - x_j}{x_i - x_j}\right) & \\ \quad + \sum_{j=i+1}^{n} \frac{x_j - x_{j-1}}{x_j - x_i} \left(1 - \frac{x_j - x_{j-1}}{x_j - x_i}\right) & \text{if } 2 \leq i \leq n-1, \\ \displaystyle\sum_{j=1}^{n-1} \frac{x_{j+1} - x_j}{x_n - x_j} \left(1 - \frac{x_{j+1} - x_j}{x_n - x_j}\right) & \text{if } i = n. \end{cases} \tag{12}$$

By the law of the large numbers, the expectations of the height function at the data points obtained in Theorem 5 correspond to the limit of the average of the heights over a forest of isolation trees, as produced by the isolation forest algorithm. The expression for the variances can be used to bound the number of trees that are necessary to approximate the average heights with prescribed confidence, although the computational cost of evaluating the variances is at least that of evaluating the expectations directly. Nonetheless, we also have the simple bound $\mathbb{V}[h_{\mathcal{T}}(x_i)] \leq \mathbb{E}[h_{\mathcal{T}}(x_i)]$, and by using an integral-series comparison with the function $x \mapsto 1/|x - x_i|$, each sum in (11) may be upper bounded by a logarithmic term which leads to the upper bound:

$$\mathbb{V}[h_{\mathcal{T}}(x_i)] \leq \begin{cases} 1 + \log \dfrac{x_n - x_1}{x_2 - x_1} & \text{if } i = 1, \\ 2 + \log \dfrac{x_n - x_i}{x_{i+1} - x_i} + \log \dfrac{x_i - x_1}{x_i - x_{i-1}} & \text{if } 2 \leq i \leq n-1, \\ 1 + \log \dfrac{x_n - x_1}{x_n - x_{n-1}} & \text{if } i = n. \end{cases}$$

The expressions obtained in Theorem 5 serve as a basis for the non asymptotic analysis that we develop in Section 5 and for the study of the isolation tree heights in the large sample regime, whereby the number of samples tends to infinity, that we expose in Section 6 in the cases of a random design and of a sequence of fixed designs.

Building upon Theorem 5, the following Proposition gives the value of the average height $\mathbb{E}[h_{\mathcal{T}}(x)]$ at any point x, while Theorem 5 gives the values of the average heights at the data points only. We state the result in a more general setting than that of Theorem 5 for further use, where the data points of $\mathcal{D}_n$ are arranged as a d-dimensional grid. When $d \geq 2$, we apply Algorithm 1 to $\mathcal{D}_n$ with the modification that in step 2 the component j along which the cell is partitioned is drawn uniformly among the components of the affine span of the set of points within that cell, meaning that if all points are identical in some dimension, then this dimension is not selected for partitioning. This is necessary in this case due to the arrangement of the points as a grid parallel to the coordinate axes. We note that when a cell is partitioned along some component j, the set of distinct values of the j'-th coordinate of the points within that cell changes only for $j' = j$. This implies that when $d \geq 2$, the isolation factorizes into several independent univariate isolations. Using this together with Theorem 5, we deduce the analytical expressions of the average heights at each point in the convex hull of $\mathcal{D}_n$, denoted by $\mathrm{Conv}(\mathcal{D}_n)$.

Proposition 6. *Let $x_{\ell,1} < x_{\ell,2} < \cdots < x_{\ell,n}$, for $\ell \in \{1, \dots, d\}$, be d collections of points that are arranged in strictly increasing order. Let $x_{\mathbf{i}} = (x_{1,i_1}, x_{2,i_2}, \dots, x_{d,i_d})$ for $\mathbf{i} = (i_1, \dots, i_d) \in \{1, \dots, n\}^d$, and let $\mathcal{D}_n = \{x_{\mathbf{i}} : \mathbf{i} \in \{1, \dots, n\}^d\}$. Let $(\mathcal{T}, \pi_{\mathcal{T}})$ denote a random isolation tree grown from $\mathcal{D}_n$.*

1. *Let $x \in \mathrm{Conv}(\mathcal{D}_n)$ be a point that belongs to the convex hull of $\mathcal{D}_n$. Let $\mathbf{i}(x) = (i_1(x), \dots, i_d(x))$ where $i_j(x) = 1 + \lfloor (n-1)(x^{(j)} - x_{j,1})/(x_{j,n} - x_{j,1}) \rfloor$. Then $\mathbb{E}[h_{\mathcal{T}}(x)]$ is a convex combination of $\{\mathbb{E}[h_{\mathcal{T}}(x_{\mathbf{i}}) : \mathbf{i} \in \mathbf{I}(x)\}$ where $\mathbf{I}(x) = \{\mathbf{i}(x) + \delta : \delta \in \{0,1\}^d\}$ and we have*
$$\mathbb{E}[h_{\mathcal{T}}(x)] = \sum_{\mathbf{i} \in \mathbf{I}(x)} \alpha_{\mathbf{i}} \mathbb{E}[h_{\mathcal{T}}(x_{\mathbf{i}})], \tag{13}$$
with
$$\alpha_{\mathbf{i}} = \prod_{j=1}^{d} \left(\frac{x^{(j)} - x_{j,i_j(x)}}{w^j} \delta_j(\mathbf{i}) + \left(1 - \frac{x^{(j)} - x_{j,i_j(x)}}{w^j}\right) (1 - \delta_j(\mathbf{i})) \right), \tag{14}$$
and with $\delta_j(\mathbf{i}) = i_j - i_j(x)$, for $\mathbf{i} \in \mathbf{I}(x)$.

2. *Let $x \in \mathbb{R}^d$. Then $\mathbb{E}[h_{\mathcal{T}}(x)] = \mathbb{E}[h_{\mathcal{T}}(x^\dagger)]$, where $x^\dagger$ is the projection of x onto $\mathrm{Conv}(\mathcal{D}_n)$.*

Thus in particular in dimension $d = 1$, by Proposition 6, the average heigth at any $x \in \mathbb{R}$ is obtained by linear interpolation of the average heights at the data points $x_1, \dots, x_n$, the expressions of whose are given by Theorem 5. This allows to compute the (theoretical) anomaly score that is produced by the isolation forest trained on $x_1, \dots, x_n$ for a new data point x. By Proposition 6, we also see that the average of the height function is continuous over $\mathbb{R}^d$, although for each isolation tree $(\mathcal{T}, \pi_{\mathcal{T}}) \in \mathbb{T}$, the height function $x \mapsto h_{\mathcal{T}}(x)$ is not since it is piecewise constant. That said, continuity of the average height function is not preserved at the limit where the number of samples tends to infinity, as we prove in Section 6.

5 Anomaly detection over a finite sample

In this section we analyze the performance of the isolation forest method as an outlier detector within a finite sample using the expressions of the average heights obtained in Theorem 5. The setting is that of n ordered distinct and fixed (deterministic) points $x_1 < x_2 < \cdots < x_n$ and we assume, without loss of generality, that $x_1 = 0$ and $x_n = 1$. We consider several configurations of points over $[0, 1]$ starting with configurations composed of one outlier and a dense cluster, and next with a configuration composed of one outlier located between two dense clusters. The proofs for the bounds stated in equations (15)—(18) are given in Section A.3.

5.1 One outlier and a dense cluster

General configuration Fix $\epsilon \in (0, 1)$ and set $x_2 = 1 - \epsilon$. Thus the data is composed of one isolated point $x_1 = 0$, which is considered as an anomaly, and of a dense cluster of $n - 1$ points $\{x_2, \dots, x_n\}$ which extends over the interval $[1 - \epsilon, 1]$. Note that the only assumption on the points forming the dense cluster is that their range is the interval $[1 - \epsilon, 1]$; in particular, the distribution of the points inside the interval is unspecified. By applying Theorem 5, we obtain that
$$\mathbb{E}[h_{\mathcal{T}}(x_1)] \leq 1 + \frac{\epsilon}{1 - \epsilon} \quad \text{and} \quad \mathbb{E}[h_{\mathcal{T}}(x_i)] \geq 2 - \epsilon, \quad \text{for } i \geq 2. \tag{15}$$
Therefore, whenever ϵ is small enough (in detail, whenever $\epsilon < c := (3 - \sqrt{5})/2$), we have $\mathbb{E}[h_{\mathcal{T}}(x_1)] < \mathbb{E}[h_{\mathcal{T}}(x_i)]$ for all $i \geq 2$. Consequently, when the configuration is such that $\epsilon < c$, there exists a threshold $\tau > 0$ such that $\mathbb{E}[h_{\mathcal{T}}(x_1)] < \tau$ and $\mathbb{E}[h_{\mathcal{T}}(x_i)] > \tau$ for all $i \geq 2$, which implies that x_1 is correctly detected as the only anomaly among the n points by thresholding the average heights at τ. That being said, the range of threshold values that yield perfect

anomaly detection, in the sense that x_1 is detected as the only anomaly among $\{x_1, \dots, x_n\}$, vary significantly according to the distribution of $\{x_2, \dots, x_n\}$ in $[1-\epsilon, 1]$. To illustrate this point, we define two configurations of points with $x_1 = 0$ and a dense cluster of $n-1$ points in $[1-\epsilon, 1]$ as before. In the first configuration, the points forming the dense cluster are evenly spaced in $[1-\epsilon, 1]$, while in the second configuration the points follow a geometric pattern (while still remaining in $[1-\epsilon, 1]$).

Configuration 1: uniform dense cluster. Let $\epsilon \in (0, 1)$. We set $x_1 = 0$ and $x_i = 1 - \epsilon + (i-2)\frac{\epsilon}{n-2}$, for $i = 2, \dots, n$. Using Theorem 5, we deduce that

$$\mathbb{E}[h_{\mathcal{T}}(x_1)] \leq 1 + \frac{\epsilon}{1-\epsilon} \quad \text{and} \quad \mathbb{E}[h_{\mathcal{T}}(x_i)] \geq \log(n-1), \quad \text{for } i \geq 2. \tag{16}$$

Consequently, x_1 is correctly detected as the only anomaly for any choice of threshold within an interval of length at least $\log(n-1) - 2$ when $\epsilon < c$, using the facts that $c < 1/2$ and that $\epsilon/(1-\epsilon) < 1$ when $\epsilon < 1/2$.

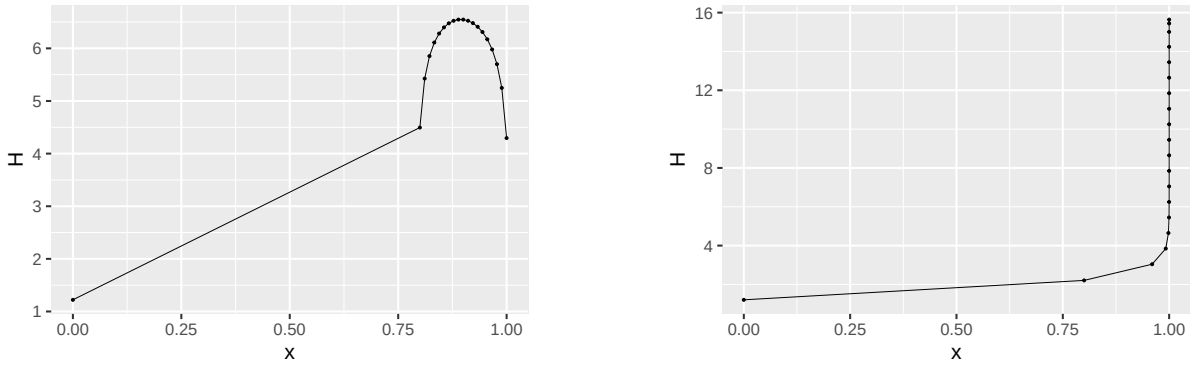
Configuration 2: geometric dense cluster. Given $\epsilon \in (0, 1)$, the configuration of points $\{x_1, \dots, x_n\}$ is defined by the recursion $x_1 = 0$ and $x_{j+1} = 1 - \epsilon(1 - x_j)$. The interval lengths satisfy the geometric recursion $w_{j+1} = \epsilon w_j$ with $w_1 = 1 - \epsilon$, so that $w_j = (1-\epsilon)\epsilon^{j-1}$, for $j = 1, \dots, n-1$. Let $\Delta_i = \mathbb{E}[h_{\mathcal{T}}(x_{i+1})] - \mathbb{E}[h_{\mathcal{T}}(x_i)]$, for $i = 1, \dots, n-1$. Using Theorem 5, we obtain that

$$\sup_{1 \leq i \leq n-1} |\Delta_i - 1| \leq 2\epsilon \quad \text{and} \quad -\epsilon \leq \Delta_n \leq 0. \tag{17}$$

When ϵ is small enough, all the gaps Δ_i's are positive and approximately equal with $\Delta_i \approx 1$, but Δ_n which is negative with $\Delta_n \approx -\epsilon$. Therefore, the average heights increase with x_i but for the last hop from x_{n-1} to x_n. In particular, x_1 has the smallest average height but correctly detecting x_1 as the only anomaly within $\{x_1, \dots, x_n\}$ requires a threshold that belongs to an interval of length Δ_1, which is approximately equal to 1, and this holds for any sample size n. Thus, when comparing the average heights values, $\mathbb{E}[h_{\mathcal{T}}(x_1)]$ is not significantly separated from $\{\mathbb{E}[h_{\mathcal{T}}(x_i)] : 2 \leq i \leq n\}$, while geometrically, x_1 is isolated from the cluster points $\{x_2, \dots, x_n\}$ which are all packed in the interval $[1-\epsilon, 1]$. This stands in sharp contrast with the setting of Configuration 1 and reveals an effect of scale in the isolation forest methodology. Indeed, if x_1 is removed from the data set, then x_2 becomes geometrically isolated from $\{x_3, \dots, x_n\}$ in the same way x_1 is isolated from $\{x_2, \dots, x_n\}$. Therefore, that a data point may be efficiently isolated and diagnosed as an outlier by the isolation forest method depends not only on such a point being geometrically isolated, but also on the distribution of the remaining points when looked at comparable scales.

The average heights for the two configurations are represented in Figure 3. The sample size is taken as $n = 20$. The values obtained for the average heights at the data points are comprised between 1.22 and 6.55 for configuration 1 and between 1.21 and 15.44 for configuration 2. In both cases, the smallest average height corresponds to x_1, illustrating the fact that x_1 can be detected as the only anomaly in both configurations by using a suitable thresholding of the average heights. We also note that the difference $\inf_{2 \leq i \leq n} \mathbb{E}[h_{\mathcal{T}}(x_i)] - \mathbb{E}[h_{\mathcal{T}}(x_1)]$ is significantly larger in configuration 1 than in configuration 2, as well as the (almost) constant difference in average height between two consecutive points in configuration 2.

To interpret the average heights as an anomaly score comprised between 0 and 1, Liu et al. (2008) introduce the scoring function $x \mapsto s(x)$ as defined in (4) and propose to operate the detection by thresholding the scoring function at a fixed threshold t, taken as either $t = 0.5$ or $t = 0.6$. In fact, the average heights obtained for these two configurations suggest that a data dependent choice of threshold may be required for the perfect detection of x_1 as the only anomaly, or at least that the threshold depends on the sample size. Indeed, at the population



(a) Configuration 1

(b) Configuration 2

Figure 3: Average heights for $n = 20$ points with one anomaly at 0 and a dense cluster of points in $[0.8, 1]$ arranged in uniform configuration (Configuration 1) and in geometric configuration (Configuration 2).

level (with respect to the forest), we have $s(x) = \exp\left(-\frac{\log(2)}{c(n)}\mathbb{E}[h_{\mathcal{T}}(x)]\right)$ and so for any $t \in (0,1)$, the inequality $s(x) > t$ is equivalent to the inequality $\mathbb{E}[h_{\mathcal{T}}] < \tau_t$ with $\tau_t = \frac{\log(1/t)}{\log(2)}c(n)$. For instance with $n = 20$, which corresponds to the setting of Figure 3, we have $c(n) \approx 5.20$ which gives $\tau_{0.5} \approx 5.20$ and $\tau_{0.6} \approx 3.83$. Comparing the heights represented in Figure 3 with these thresholds, we obtain that with $t = 0.5$, the detected anomalies are $\{x_1, x_2, x_n\}$ for configuration 1 and $\{x_1, \dots, x_5\}$ for configuration 2, while with $t = 0.6$, only x_1 is detected as an anomaly in configuration 1 and $\{x_1, x_2, x_3\}$ are detected as anomalies in configuration 2. More generally, using the fact that $c(n) \sim 2\log(n)$ as $n \to \infty$ together with (16) (for configuration 1) and (15) and (17) combined (for configuration 2), we see that the property that there exists some integer n_0 and some fixed $t \in (0,1)$ such that $s(x_1) > t$ and $\inf_{i \geq 2} s(x_i) < t$ holds for all $n \geq n_0$ is valid for configuration 1 but not for configuration 2. For this property to be valid when the points are arranged in configuration 2, the thresholding of the scoring function must actually depend on n.

5.2 One outlier between two dense clusters

Here we consider a configuration of n distinct points $x_1 < \cdots < x_n$ where, for some integer $3 < k < n-2$, $x_k = \frac{1}{2}$, and where $\{x_1, \dots, x_{k-1}\}$ and $\{x_{k+1}, \dots, x_n\}$ extend over the intervals $[0, \epsilon]$ and $[1-\epsilon, 1]$, respectively, where $\epsilon \in (0, 1/4)$ is fixed. Thus geometrically, x_k is considered as an anomaly located between two dense clusters. Using Theorem 5, we obtain that

$$\mathbb{E}[h_{\mathcal{T}}(x_k)] \leq 2 + 8\epsilon \quad \text{and} \quad \mathbb{E}[h_{\mathcal{T}}(x_i)] \geq \frac{5}{2} - 3\epsilon, \quad \text{for any } i \neq k. \tag{18}$$

Therefore, when ϵ is taken small enough, we have $\mathbb{E}[h_{\mathcal{T}}(x_k)] < \mathbb{E}[h_{\mathcal{T}}(x_i)]$ for all $i \neq k$ and so x_k can be correctly detected as the only anomaly among $\{x_1, \dots, x_n\}$. As for the two configurations considered in the previous section, the difference $\inf_{2 \leq i \leq n} \mathbb{E}[h_{\mathcal{T}}(x_i)] - \mathbb{E}[h_{\mathcal{T}}(x_1)]$ may vary significantly (from being constant with n to being on the order of $\log(n)$) depending on the distribution of the points in the two clusters.

6 Asymptotic analysis

In this section we study the average height function as the sample size tends to infinity, first in a random design (Section 6.1) and next in a sequence of fixed designs (Section 6.2)

6.1 Random design

We consider an IID random sample $X_1, \dots, X_n$ drawn from a distribution F with probability density function f on $\mathbb{R}$. The isolation forest method is applied to the sample, and we focus on the average height function of the forest trees as the number of samples tends to infinity. Let us point out that we do not consider a combined asymptotic regime with a finite number of trees tending to infinity at the same time as the number of samples. Instead, we build upon Theorem 5 and let $n \to \infty$. This amounts at considering the asymptotic of the scoring of an infinite forest of trees as n goes to inifinity. Arguably this is justified since in practice the number of trees in an isolation forest can be chosen as large as desired and is only limited by a time or a computational budget.

Let $(\mathcal{T}, \pi_{\mathcal{T}})$ be a random isolation tree grown from the random sample $\{X_1, \dots, X_n\}$, where the draws of the split points, as described in Algorithm 1, are independent from the sample. Denote by $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ the ordered sample. Let $\bar{H}_i = \mathbb{E}\left[h_{\mathcal{T}}(X_{(i)}) \,|\, X_1, \dots, X_n\right]$ be the conditional expectation of the height of the leaf of $\mathcal{T}$ whose cell contains $X_{(i)}$ given the sample.

Then, by Theorem 5 we have, almost surely,

$$\bar{H}_i = \begin{cases} \displaystyle\sum_{j=2}^{n} \frac{X_{(j)} - X_{(j-1)}}{X_{(j)} - X_{(1)}} & \text{if } i = 1, \\ \displaystyle\sum_{j=1}^{i-1} \frac{X_{(j+1)} - X_{(j)}}{X_{(i)} - X_{(j)}} + \sum_{j=i+1}^{n} \frac{X_{(j)} - X_{(j-1)}}{X_{(j)} - X_{(i)}} & \text{if } 2 \leq i \leq n-1, \\ \displaystyle\sum_{j=1}^{n-1} \frac{X_{(j+1)} - X_{(j)}}{X_{(n)} - X_{(j)}} & \text{if } i = n. \end{cases} \tag{19}$$

Among the possible ways of formulating the convergence of the heights, we find it convenient in dimension one to make use of the quantile function $G = F^{-1}$, defined by $G(p) = \inf\{x : F(x) \geq p\}$ for $0 \leq p \leq 1$. In the proofs of our results, a control on the tail probabilities of $X \sim F$ is needed when the support of F is unbounded, and so we make the assumption that X is sub-exponential, meaning that X is integrable and that there exists non-negative parameters (σ, b) such that $\mathbb{E}\left[\exp\left(\lambda(X - \mathbb{E}[X])\right)\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for all $|\lambda| < \frac{1}{b}$, in which case X satisfies the following tail bound:

$$\mathbb{P}\left(X \geq \mathbb{E}[X] + u\right) \leq \begin{cases} \exp\left(-\frac{u^2}{2\sigma^2}\right) & \text{if } 0 \leq u \leq \frac{\sigma^2}{b} \\ \exp\left(-\frac{u}{2b}\right) & \text{if } u > \frac{\sigma^2}{b}; \end{cases} \tag{20}$$

see for instance Wainwright (2019, Definition 2.7 & Proposition 2.9).

We start with a pointwise convergence result using mild local regularity assumptions on the underlying density f. We cover the cases of a fixed quantile of order $p \in (0,1)$, and those cases where the support of the density is bounded from below or from above.

Theorem 7. *Assume that the distribution F is sub-exponential.*

(i) Let $p \in (0,1)$ and let $x_p = F^{-1}(p)$. Assume that f is continuously differentiable in an open neighborhood of x_p and that $f(x_p) > 0$. Let $(i_n(p))_{(n \geq 1)}$ be a sequence of integers such that $1 \leq i_n(p) \leq n$ for all $n \geq 1$ and such that $\frac{i_n(p)}{n} \to p$ as $n \to \infty$. Then

$$\frac{1}{\log n} \bar{H}_{i_n(p)} \to 2, \quad \text{almost surely as } n \to \infty.$$

(ii) Let $a = \inf\{x : F(x) > 0\}$. Assume that $a > -\infty$, that f is continuously differentiable over $(a, a+\epsilon)$ for some $\epsilon > 0$ and with a right-continuous right-derivative at a, and that $f(a) > 0$. Then

$$\frac{1}{\log n} \bar{H}_1 \to 1, \quad \text{almost surely as } n \to \infty.$$

(iii) *Let* $b = \sup\{x : F(x) < 1\}$. *Assume that* $b < \infty$, *that* f *is continuously differentiable over* $(b-\epsilon, b)$ *for some* $\epsilon > 0$ *and with a left-continuous left-derivative at* b, *and that* $f(b) > 0$. *Then*

$$\frac{1}{\log n}\bar{H}_n \to 1, \quad \textit{almost surely as } n \to \infty.$$

With a uniform version of the regularity assumption made on f, we establish a uniform convergence result of the tree heights over a closed interval of quantiles.

Theorem 8. *Assume that the distribution* F *is sub-exponential. Let* $0 < p_1 < p_2 < 1$, *and let* $x_{p_1} = F^{-1}(p_1)$ *and* $x_{p_2} = F^{-1}(p_2)$. *Assume that* f *is continuously differentiable on an open neighborhood of* $[x_{p_1}, x_{p_2}]$ *and bounded away from* 0 *on* $[x_{p_1}, x_{p_2}]$. *Then*

$$\sup_{p_1 \le p \le p_2} \left| \frac{1}{\log n}\bar{H}_{\lfloor pn \rfloor} - 2 \right| \to 0, \quad \textit{almost surely as } n \to \infty.$$

By Proposition 6, the value of $\mathbb{E}[h_{\mathcal{T}}(x)|X_1, \dots, X_n]$ at any x is obtained by linear interpolation of $\{(X_{(i)}, \bar{H}_i) : i = 1, \dots, n\}$ when x is in the range of the data, and to either $\bar{H}_1$ or $\bar{H}_n$ when $x < X_{(1)}$ or $x > X_{(n)}$ respectively. Thus, for any $x \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{E}\left[h_{\mathcal{T}}(x)|X_1, \dots, X_n\right] \\
&\quad = \bar{H}_1 \mathbf{1}\{x < X_{(1)}\} + \bar{H}_n \mathbf{1}\{x \ge X_{(n)}\} \\
&\qquad + \sum_{i=1}^{n-1} \left[\left(1 - \frac{x - X_{(i)}}{X_{(i+1)} - X_{(i)}} \right) \bar{H}_i + \frac{x - X_{(i)}}{X_{(i+1)} - X_{(i)}} \bar{H}_{i+1} \right] \\
&\qquad\qquad \times \mathbf{1}\{X_{(i)} \le x < X_{(i+1)}\}.
\end{aligned}$$

Combining this with Theorem 7 and Theorem 8 immediately leads to the following Corollary. We denote by $\mathcal{S} = \overline{\{x \in \mathbb{R} : f(x) > 0\}}$ the support of the density f. The boundary and interior of $\mathcal{S}$ are denoted by $\partial\mathcal{S}$ and $\mathring{\mathcal{S}}$ respectively. For simplicity, we assume that $\partial\mathcal{S}$ contains at most two points. With a bit of extra work, the conclusions of Corollary 9 remain valid if $\mathcal{S}$ is a disjoint union of intervals. In essence, the limit value of $\mathbb{E}[h_{\mathcal{T}}(x)|X_1, \dots, X_n]$ at any x is obtained by reproducing the steps in the proofs of Theorem 7 and Theorem 8 using the observations that belong to the same interval as x, and those latter can be identified with high enough confidence by means of statistics of the form $X_{(i+1)} - X_{(i)} \ge \epsilon_n$ given a conveniently chosen sequence (ϵ_n) that tends to 0 as $n \to \infty$. We elaborate on this setting in Theorem 11 in the context of a sequence of fixed designs. The setting and assumptions of Corollary 9 are those of Theorem 7 and Theorem 8 combined, which means that we assume that F is sub-exponential, and that f is continuously differentiable over $\mathring{\mathcal{S}}$, and if $\partial\mathcal{S} \neq \emptyset$, we assume that, at any $x \in \partial S$, $f(x) > 0$ and that f admits a right-continuous right-derivative (resp. left-continuous left derivative) if x is a left (resp. right) limit point of $\mathcal{S}$.

Corollary 9. *Let* $(X_i)_{i \ge 1}$ *be a sequence of independent random variables each with distribution* F *satisfying the assumptions above. Then*

$$\frac{1}{\log(n)}\mathbb{E}[h_{\mathcal{T}}(x) \,|\, X_1, \dots, X_n] \to \begin{cases} 1 & \textit{for any } x \in \partial\mathcal{S} \textit{ when } \partial\mathcal{S} \neq \emptyset, \\ 2 & \textit{for any } x \in \mathring{\mathcal{S}} \textit{ with } f(x) > 0, \end{cases}$$

almost surely as $n \to \infty$. *Moreover, almost surely, the convergence is uniform over any closed subset* $\mathcal{K}$ *of* $\mathbb{R}$ *included in* $\mathring{\mathcal{S}}$ *such that* $\inf\{f(x) : x \in \mathcal{K}\} > 0$, *meaning that*

$$\sup_{x \in \mathcal{K}} \left| \frac{1}{\log(n)}\mathbb{E}\left[h_{\mathcal{T}}(x)|X_1, \dots, X_n\right] - 2 \right| \to 0 \quad \textit{almost surely as } n \to \infty.$$

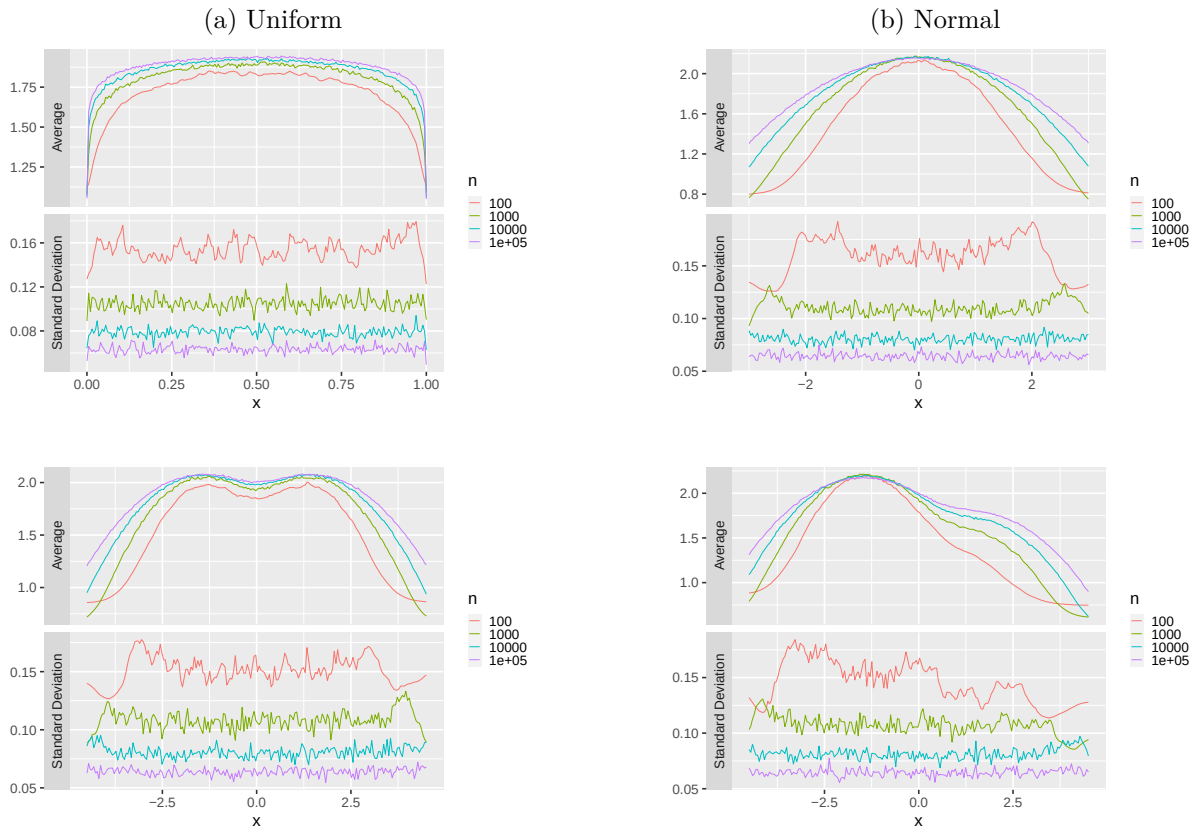


Figure 4: Numerical examples illustrating the setting of Corollary 9. Here $N = 200$ samples of sizes n^k, with $k \in \{2, 3, 4, 5\}$, were simulated according to a uniform distribution, a standard normal distribution, and mixtures of normal distributions with variances equal to 1, means equal to -1.5 and 1.5, and mixture coefficients equal to 0.5 and 0.5 (mixture 1) and 0.9 and 0.1 (mixture 2). The pointwise average and standard deviation of $\mathbb{E}[h_{\mathcal{T}}(x) \,|\, X_1, \ldots, X_n]$ were evaluated on the N samples at 200 equally spaced points x that extend over $[0, 1]$ (uniform case), over $[-3, 3]$ (standard normal case) and over $[-4.5, 4.5]$ (mixture cases).

Thus in the large sample regime the isolation forest methodology operates as a detector of the support of the underlying distribution. For instance, consider an isolation forest anomaly detector trained on a number n of data $x_{(1)} < \cdots < x_{(n)}$ drawn from a density f of class C^1, supported on a closed interval $[a, b]$ and bounded away from 0 over $[a, b]$. Then when n is large enough, $\mathbb{E}[h_{\mathcal{T}}(x) \,|\, X_1, \ldots, X_n]/\log(n)$ is uniformly close to 2 over any closed interval included in (a, b), while $\mathbb{E}[h_{\mathcal{T}}(x) \,|\, X_1, \ldots, X_n]/\log(n)$ is close to 1 for any $x \leq a$ and any $x \geq b$. In other words, any upper level set of $\mathbb{E}[h_{\mathcal{T}}(x) \,|\, X_1, \ldots, X_n]$ of the form $\{x \in \mathbb{R} : \mathbb{E}[h_{\mathcal{T}}(x) \,|\, X_1, \ldots, X_n] \geq \tau \log(n)\}$ with $1 < \tau < 2$ is a consistent estimator of $[a, b]$.

Numerical examples are given in Figure 4. We consider the cases of a uniform distribution over $[0, 1]$, of a standard normal distribution, and of two mixtures of two normal distributions with variances equal to 1 and means equal to -1.5 and 1.5 respectively. The mixture coefficients are taken as 0.5 and 0.5 in the first mixture, and as 0.9 and 0.1 in the second mixture. For each sample size n of the form $n = 10^k$, with $k \in \{2, 3, 4, 5\}$, we run Monte Carlo simulations based on $N = 200$ samples $\mathcal{D}_n$ and we estimate the pointwise average and standard deviation of the scaled average height function $x \mapsto \mathbb{E}[h_{\mathcal{T}}(x) \,|\, \mathcal{D}_n]/\log(n)$ at 200 equally spaced points x. The evaluation points extend over $[0, 1]$ in the case of the uniform distribution, over $[-3, 3]$ in the case of the standard normal distribution, and over $[-4.5, 4.5]$ in the cases of the mixtures.

These simulations numerically confirm the uniform convergence of the height function (scaled by a $\log(n)$ factor). In these simulations, that the limit function is equal to $x \mapsto 1 + \mathbf{1}\{x \in \mathcal{S}\}$ for a distribution with support $\mathcal{S}$ is apparent in the case of the uniform distribution, which illustrates the fact that in the large sample regime, scoring with an isolation forest essentially amounts at estimating $\mathcal{S}$. The convergence towards a limit value of 2 is less visible in the cases of the standard normal and of the normal mixtures, especially in the areas of low density. In fact, as we argue in the next section, the convergence occurs at a logarithm rate in the sample size, so that large sample sizes may be necessary for the convergence to be evidenced within a prescribed accuracy through simulations. We anticipate that for any $x \in \mathring{\mathcal{S}}$, $\mathbb{E}[h_{\mathcal{T}}(x)] - 2\log(n)$ converges towards a constant $c(x)$ such that $|c(x)|$ is all the more large as x is close to $\partial\mathcal{S}$ (when the boundary is non empty) or as $f(x)$ is close to 0, as the simulations suggest.

Tree height limit We conclude this section with a remark on the asymptotic effect of imposing a limit condition on the tree height during the growth of an isolation forest. Indeed following Liu et al. (2008, 2012) trees are typically built up to a height of $\log_2(n)$ in practice. More generally, let $c > 0$ be a fixed positive real number and suppose that a height limit is imposed on the isolation trees at $c\log(n)$. Then under the setting of Corollary 9, the resulting average height at any point $x \in \mathbb{R}$ is given by $\mathbb{E}[h_{\mathcal{T}}(x) \wedge c\log(n)|X_1, \dots, X_n]$. By Theorem 5, we have $\mathbb{V}[h_{\mathcal{T}}(x)|X_1, \dots, X_n] \leq \mathbb{E}[h_{\mathcal{T}}(x)|X_1, \dots, X_n]$ implying, when combined with Corollary 9, that $\mathbb{V}\left[\frac{1}{\log n} h_{\mathcal{T}}(x)|X_1, \dots, X_n\right] \to 0$ almost surely as $n \to \infty$. Using this, we deduce that

$$\frac{1}{\log(n)}\mathbb{E}[h_{\mathcal{T}}(x) \wedge c\log(n)|X_1, \dots, X_n] \to \begin{cases} 1 \wedge c & \text{for any } x \in \partial\mathcal{S} \text{ when } \partial\mathcal{S} \neq \emptyset, \\ 2 \wedge c & \text{for any } x \in \mathring{\mathcal{S}} \text{ with } f(x) > 0, \end{cases}$$

almost surely as $n \to \infty$. The choice of $c = 1/\log(2) \approx 1.446$ corresponds to a tree height limit set at $\log_2(n)$. In this case, the set of thresholds that are admissible for asymptotically detecting the support reduces from $[1, 2]$ to $[1, 1.446]$. If $c < 1$, the limit function is constant equal to c and the support is not detected at the limit. Finally, if $c > 2$, then the limit is identical to that of Corollary 9 when no height limit is imposed.

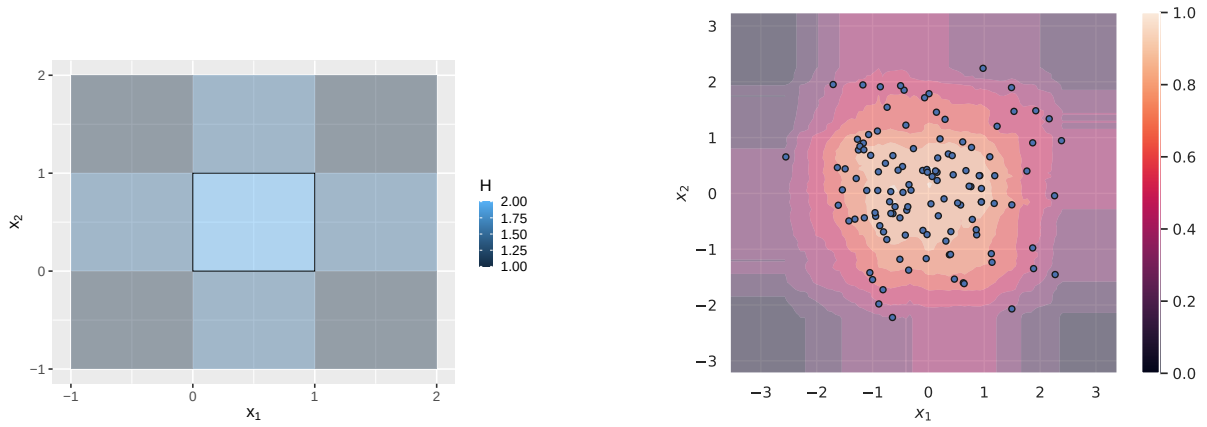
6.2 Fixed design

In this section we consider a configuration of n^d points arranged in a full regular grid over the unit cube of $\mathbb{R}^d$, and we study the asymptotic behavior of the average height function as the number of samples tends to infinity. This scenario is prototypical of the case of a random sample that would be drawn from a distribution that resembles the uniform distribution over the unit cube of $\mathbb{R}^d$, in the sense that the distribution admits a density that is bounded from below and from above by strictly positive numbers. The convergence result that we obtain (Theorem 10) also holds in the slightly more general setting of points arranged in a full irregular grid of a hyperrectangle of $\mathbb{R}^d$, as considered in Proposition 6. In fact, the main technical requirement is to be able to apply Theorem 5 by tensorization, so we only consider a regular design over the unit cube.

Let $\mathcal{D}_n$ be the set of n^d points defined by

$$\mathcal{D}_n = \left\{ x_{\mathbf{i}} = \left(\frac{i_1 - 1}{n - 1}, \dots, \frac{i_d - 1}{n - 1} \right) \in \mathbb{R}^d \,:\, \mathbf{i} = (i_1, \dots, i_d) \in \{1, \dots, n\}^d \right\}. \tag{21}$$

As in the setting of Proposition 6, for each sample size n, Algorithm 1 is applied to $\mathcal{D}_n$ with the modification that in step 2 the component j along wich the cell is partitioned is drawn uniformly among the components of the affine span of the set of points within that cell, and trees are grown until each point is isolated. Using Theorem 5 together with Proposition 6, we deduce the analytical expressions of the average heights at each point in $\mathcal{D}_n$ and we derive their (scaled) limit as n goes to infinity in Theorem 10.



(a) Regular design

(b) Normal distribution

Figure 5: Numerical example illustrating the setting of Theorem 10 showing that the scoring function admits axis-aligned discontinuities in regions outside the support of the distribution. Here are represented in subfigure (a) the limit function of Theorem 10 for the regular design (21) over $[0,1]^2$, and in subfigure (b) the scoring function (normalized between 0 and 1) that results from applying the isolation forest algorithm to a sample of size $n = 120$ drawn from a standard multivariate normal distribution over $\mathbb{R}^2$.

For points belonging to the boundary of the cube, the scaled heights are found to converge to values that depend on the locations of the points on the boundary. For any $k \in \{0, \dots, d\}$, let $\mathcal{F}_k$ be the set of k-dimensional faces of $[0,1]^d$, where a face of the cube is defined as any set of the form $[0,1]^d \cap \{x \in \mathbb{R}^d : \langle x, a \rangle = a_0\}$ for some $x \in \mathbb{R}^d$ and $a_0 \in \mathbb{R}$ such that $\langle x, a \rangle \leq a_0$ is a valid inequality for $[0,1]^d$, meaning that it is satisfied for all points in $[0,1]^d$ (see Ziegler, 1995, Definition 2.1). The elements of $\mathcal{F}_0$ are the vertices of the cube and $\mathcal{F}_d$ is the cube itself (note that, for any $k \in \{1, \dots, d\}$, the boundary of any face in $\mathcal{F}_k$ is an element of $\mathcal{F}_{k-1}$).

Theorem 10. *Let $(\mathcal{T}, \pi_{\mathcal{T}})$ be a random isolation tree defined using Algorithm 1 with $\mathcal{D}_n$ as defined in (21). Then for any $k \in \{0, 1, \dots, d\}$, any face $F_k \in \mathcal{F}_k$, and any $x \in \mathring{F}_k$, we have*

$$\frac{1}{d \log(n)} \mathbb{E}[h_{\mathcal{T}}(x)] \to 1 + \frac{k}{d}.$$

Moreover, the convergence is uniform over any closed subset contained in the interior of $[0,1]^d$.

Thus, in this fixed design scenario, the limit of the scoring function admits axis-aligned discontinuities outside the support of the data (here, the unit cube). This result explains the artifacts that are observed in practice on the scoring function that is produced by the isolation forest method and that motivated the introduction of the extended isolation forest variant by Hariri et al. (2021). A numerical example is provided in Figure 5, where we used the scikit-learn toolkit (Pedregosa et al., 2011) for the isolation forest simulation on normal data.

We also note that the values of the scaled average heights at the limit are equal to 2 at points belonging to the interior of the cube, and that they extend over $[1, 1+(d-1)/d]$ at points exterior to the cube. So recovering the support at the limit requires a threshold τ comprised between $1+(d-1)/d$ and 2, thus within a range of length $1/d$ which goes to 0 as $d \to \infty$. This suggests a decrease in performance in high dimension of anomaly detection by an isolation forest in a large sample regime where performance is conceived as support recovery.

In the case of the regular design of (21), the expressions of the average heights obtained in Theorem 5 and that we use to prove Theorem 10 simplify to sums of several harmonic numbers $\mathcal{H}_\ell$, where $\mathcal{H}_\ell = \sum_{k=1}^{\ell} \frac{1}{k}$, for any integer $\ell \geq 1$. Using the asymptotic expansion $\mathcal{H}_\ell = \log(\ell) +$

$\gamma+o(1)$, where $\gamma \approx 0.5772...$ denotes the Euler-Mascheroni constant, it follows directly from (70) in the proof of Theorem 10, that $\mathbb{E}[h_{\mathcal{T}}(x)] - 2d\log(n)$ converges to $2d\gamma + \sum_{i=1}^{d} \log(x_i(1-x_i))$ at any $x = (x_1, \dots, x_d)$ belonging to the interior of the unit cube of $\mathbb{R}^d$. Thus, the convergence rate is logarithmic in the sample size and the limit value depends on x. In particular when $d = 1$, $\mathbb{E}[h_{\mathcal{T}}(x)] - 2\log(n)$ converges to $\gamma + \log(x(1-x))$ at any $x \in (0,1)$, and $|\log(x(1-x))|$ increases with the distance from x to $1/2$; compare with the plots given in Figure 4a for the case of a uniform random design where the estimated value of $|\mathbb{E}[h_{\mathcal{T}}(x)]/\log(n) - 2|$ shows a pattern increasing with $|x - 1/2|$.

6.3 Multiply connected support, the masking effect, and robustness

In this section, we consider the case of n points spread over K disjoint intervals of the real line. We consider regular designs where within each interval the points are equally spaced, arguing as in the previous section that, from the point of view of the asymptotic analysis, this setting is representative of the case of a random sample that would be drawn from a mixture distribution with components resembling the uniform distribution.

Given $K \geq 2$ an integer, we consider K disjoint intervals $\mathcal{I}_1 \leq \cdots \leq \mathcal{I}_K$ of the real line. For each $k \in \{1, \dots, K\}$, we denote by $L_k \in \mathbb{R}$ the length of $\mathcal{I}_k$, and we let $\delta_k > 0$ be the gap between $\mathcal{I}_k$ and $\mathcal{I}_{k+1}$, for $k \in \{1, \dots, K-1\}$. We design a configuration $x_1 < x_2 < \cdots < x_n$ of $n = n_1 + \cdots + n_K$ points, where n_k denotes the number of points that belong to $\mathcal{I}_k$, for each $k \in \{1, \dots, K\}$, that we scale over the unit interval by requiring that $\sum_{k=1}^{K} L_k + \sum_{k=1}^{K-1} \delta_k = 1$. Thus the points are defined in each interval $\mathcal{I}_k$, for $k \in \{1, \dots, K\}$, by

$$x_i = \sum_{\ell=1}^{k-1}(L_\ell + \delta_\ell) + \frac{i - 1 - \sum_{\ell=1}^{k-1} n_\ell}{n_k - 1}, \quad \text{for } i = \sum_{\ell=1}^{k-1} n_\ell + 1, \dots, \sum_{\ell=1}^{k} n_\ell, \tag{22}$$

where we use the convention that a sum over an empty range of integers is equal to 0. Then we have $\mathcal{I}_1 = [x_1, x_{n_1}]$ and $\mathcal{I}_k = [x_{n_1+\cdots+n_{k-1}+1}, x_{n_1+\cdots+n_k}]$ for $k \in \{2, \dots, K\}$, and $\delta_k = x_{k'+1} - x_{k'}$ with $k' = \sum_{\ell=1}^{k} n_\ell$, for $k \in \{1, \dots, K-1\}$. We consider first a dense asymptotic regime whereby

$$\frac{n_k}{n} \to \alpha_k > 0 \quad \text{as } n \to \infty \text{ for each } k \in \{1, \dots, K\}. \tag{23}$$

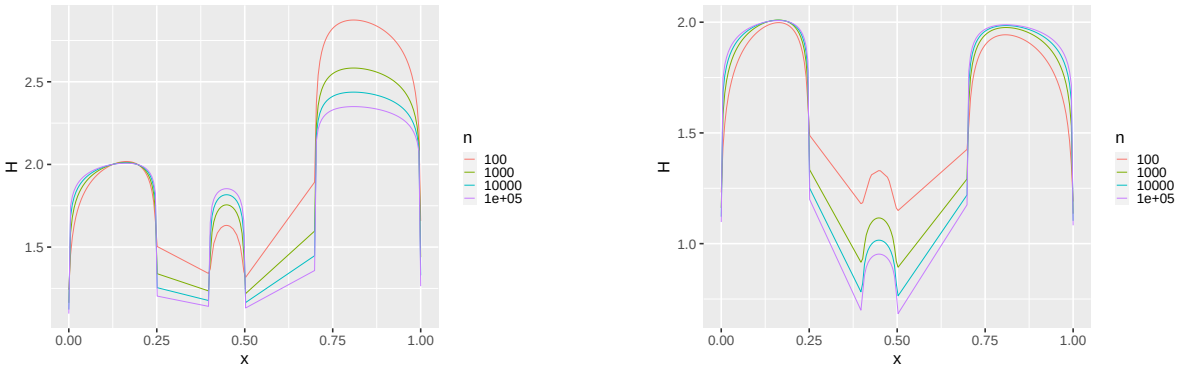
Theorem 11. *Let $\mathcal{D}_n = \{x_1 < \cdots < x_n\}$ be a configuration of n points as defined in (22). Let $(\mathcal{T}, \pi_{\mathcal{T}})$ be a random isolation tree defined using Algorithm 1 with $\mathcal{D}_n$. Then for any $x \in \mathbb{R}$, in the dense asymptotic regime (23), we have*

$$\frac{1}{\log(n)}\mathbb{E}[h_{\mathcal{T}}(x)] \to \begin{cases} 2 & \text{if } x \in \cup_{k=1}^{K} \mathring{\mathcal{I}}_k, \\ 1 & \text{otherwise.} \end{cases}$$

Moreover, the convergence is uniform over any closed subset of $\mathbb{R}$ included in $\cup_{k=1}^{K} \mathring{\mathcal{I}}_k$.

From Theorem 11, it follows that in the dense regime the isolation forest method also detects the support of the underlying distribution in the case where the support is composed of multiple connected components. Interestingly, the function obtained at the limit does not depend on the geometry of the support, in the sense that it admits only two distinct values (either 1 or 2) and that this holds for any choice of interval lengths and gaps. Therefore in the context of anomaly detection, Theorem 11 implies that if one of these components should be considered as anomalous, then points originating from this component will not be detected as anomalies, a phenomenom known as the *masking effect*. A numerical example is provided in Figure 6a.

We also note that the convergence stated in Theorem 11 does not depend on the asymptotic proportions of the components as defined in (23). In fact, as may be seen from the proof, any component with a number of points n_k satisfying $\log(n_k)/\log(n) \to 1$ will be detected as being



(a) Dense regime

(b) Sparse regime

Figure 6: Numerical examples illustrating the masking effect (a) and the robustness to training under the presence of anomalies (b). Here, samples of sizes n^k, with $k \in \{1, \dots, k\}$, are generated according to (22) with $K = 3$ intervals set as $\mathcal{I}_1 = [0, 0.25]$, $\mathcal{I}_2 = [0.4, 0.5]$ and $\mathcal{I}_3 = [0.7, 1]$. In the dense regime, the asymptotic proportions are taken as $\alpha_1 = 0.5$, $\alpha_2 = 0.1$ and $\alpha_3 = 0.4$. In the sparse regime, we set $\alpha_1 = \alpha_3 = 0.5$ and $\log(n_2)/\log(n) \to 0.35$. The values of the scaled average heights for points in the second interval converge to 2 in the dense regime, illustrating the masking effect, should the second component be considered as anomalous, while they reach a value smaller than 1 (which is the minimal asymptotic threshold for support recovery) in the sparse regime, illustrating the robustness to contamination during training.

non abnormal asymptotically. Elaborating a bit on the topic, if we consider a regime in which a component, say the k^{th} on the interval $\mathcal{I}_k$, is such that $n_k \to \infty$ and that $\log(n_k)/\log(n) \to \kappa$, for some $\kappa \in [0, 1]$, then we obtain that $\mathbb{E}[h_{\mathcal{T}}(x)]/\log(n) \to 2\kappa$ for any $x \in \mathring{\mathcal{I}}_k$ and $\mathbb{E}[h_{\mathcal{T}}(x)]\log(n) \to \kappa$ for any $x \in \partial\mathcal{I}_k$. In particular, if $\kappa \leq 1/2$, then the scaled average height at any $x \in \mathcal{I}_k$ is smaller than 1 at the limit, which is the limit value of the scaled average height for points not belonging to the support in the dense regime. More generally, if a dectector is conceived by thresholding $\mathbb{E}[h_{\mathcal{T}}(x)]/\log(n)$ at some threshold $\tau \in (1, 2)$, then points belonging to $\mathcal{I}_k$ will be detected as anomalies whenever n_k satisfies $\limsup \log(n_k)/\log(n) \leq \tau/2$. From this, we conclude that the training of an isolation forest is robust to the presence of anomalies in the training set provided anomalies aggregate in sparse clusters, in the sense that the proportion of abnormal data decays at least at the rate of $n^{-(1-\tau/2)}$ for $\tau \in (1, 2)$. A numerical example is given in Figure 6b.

The training of an isolation forest is also robust to contamination by sparse anomalies even if abnormal data do not aggregate in sparse clusters, as we illustrate by the following example over the unit interval. Fix $a \in (1/2, 1)$ and $\nu \in (0, 1)$, and let $\alpha = (1/a)/n^{1-\nu}$. Consider a configuration of n points where $n_1 := \lfloor n\alpha a \rfloor$ points extend evenly over $[0, a]$, n_1 points extend evenly over $[1-a, 1]$, and the remaining points are positioned evenly in $(a, 1-a)$, which is representative of a mixture distribution of the form $(1-\alpha)\mathcal{U}([a, 1-a]) + \alpha\mathcal{U}([0, 1])$, composed of a main component generating the normal data over $[a, 1-a]$ contaminated by anomalies that extend over $[0, 1]$. Under the asymptotic regime considered here, we have $\alpha \to 0$ and $\log(n_1)/\log(n) \to \nu$ as $n \to \infty$. By using Theorem 5 and arguing as in the proof of Theorem 11, we find that $\limsup_{n\to\infty} \mathbb{E}[h_{\mathcal{T}}(x)]/\log(n) \leq 3\nu$ for any $x \in [0, a) \cup (1-a, 1]$, that $\mathbb{E}[h_{\mathcal{T}}(x)]/\log(n) \to 1+\nu$ if $x = a$ or $x = 1-a$, and $\mathbb{E}[h_{\mathcal{T}}(x)]/\log(n) \to 2$ for any $x \in (a, 1-a)$, with convergence being uniform over any closed subset of $[0, 1]$ included in either $(0, a)$, or $(a, 1-a)$ or $(1-a, 1)$. Since $3\nu \vee (1+\nu) < 2$ when $\nu < 2/3$, with a choice of threshold τ satisfying $3\nu \vee (1+\nu) < \tau < 2$, the thresholding map $x \mapsto \mathbf{1}\{x \in [0, 1] : \mathbb{E}[h_{\mathcal{T}}(x) \leq \tau\log(n)]\}$ correctly recovers asymptotically the support $[a, 1-a]$ of the normal data. Therefore the training of an isolation forest is robust to sparse contamination by anomalies (where performance in an

asymptotic setting is apprehended as support recovery, as in the previous section).

7 Discussion

We discuss first a simple variant to the original isolation forest algorithm of Liu et al. (2008, 2012) in dimension 1 that leads to a connection with the Hilbert density estimate introduced in Devroye and Krzyżak (1999). Next we comment on subsampling and on the normalization factor used in the definition of the scoring function of an isolation forest. Then we conclude with multi-dimensional considerations.

Weighted path lengths We consider the setting of Theorem 5 where we are given n ordered points $x_1 < \cdots < x_n$ of $\mathbb{R}$. Given an isolation tree $(\mathcal{T}, \pi_{\mathcal{T}})$, the path length $\mathbb{E}[h_{\mathcal{T}}(x_i)]$ at each data point x_i gives the number of recursive partitioning operations that are needed to isolate x_i from the remaining points using $(\mathcal{T}, \pi_{\mathcal{T}})$. A simple variant consists in weighting the edges of $\mathcal{T}$ to produce weighted path lengths as basis elements to the definition of the scoring function. Among the variety of weighting mechanisms that may be envisioned, we consider the following one where the weights are in one-to-one correspondance with the intervals $I_1, \ldots, I_{n-1}$. Given a set of positive weights $\{\alpha_i \,:\, 1 \leq i \leq n-1\}$, each edge connecting an internal node $v \in \mathcal{T}^\circ$ to one of its children $v\eta$, $\eta \in \{0,1\}$ is weighted by $\alpha_{i(v)}$, where $i(v)$ is such that the value at which $\pi_{\mathcal{T}}(v)$ is partitioned belongs to $I_{i(v)}$. We denote by $(\bar{\mathcal{T}}, \pi_{\bar{\mathcal{T}}})$ an isolation tree weighted using this assignment. This changes the depth $D_i = \sum_{j=1}^n A_{ji}$ of I_i on the binary search tree into $\bar{D}_i = \sum_{j=1}^n \alpha_j A_{ji}$, and proceeding as in the proof of Theorem 5, this leads to

$$\mathbb{E}[h_{\bar{\mathcal{T}}}(x_i)] = \begin{cases} \displaystyle\sum_{j=2}^{n} \frac{\alpha_j(x_j - x_{j-1})}{x_j - x_1} & \text{if } i = 1, \\ \displaystyle\sum_{j=1}^{i-1} \frac{\alpha_j(x_{j+1} - x_j)}{x_i - x_j} + \sum_{j=i+1}^{n} \frac{\alpha_j(x_j - x_{j-1})}{x_j - x_i} & \text{if } 2 \leq i \leq n-1, \\ \displaystyle\sum_{j=1}^{n-1} \frac{\alpha_j(x_{j+1} - x_j)}{x_n - x_j} & \text{if } i = n. \end{cases} \tag{24}$$

When $\alpha_i = 1$ for all $1 \leq i \leq n-1$, we recover the unweighted case and (24) reduces to (11). Interestingly, when the weights are taken as the reciprocals of the interval lengths, so that $\alpha_i = w_i^{-1}$ for any $1 \leq i \leq n-1$, the numerators in the sums in (24) are all equal to one and this yields

$$\mathbb{E}[h_{\bar{\mathcal{T}}}(x_i)] = \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{|x_j - x_i|}, \quad \text{for all } 1 \leq i \leq n. \tag{25}$$

As it turns out, when the x_i's are random variables, expression (25) for $\mathbb{E}[h_{\bar{\mathcal{T}}}(x_i)]$ is that of the value at x_i of the Hilbert kernel density estimate introduced in Devroye and Krzyżak (1999), and where the estimate is defined on all the data but x_i. This is a one-of-a-kind kernel density estimate since it does not have a bandwidth parameter, so it automatically scales with the sample size. The Hilbert name was coined after the Hilbert transform and it is shown to be weakly consistent in Devroye and Krzyżak (1999), following previous work in the regression setting (Devroye et al., 1998). On the other hand, the Hilbert density estimate is not strongly consistent, has poor rate of convergence as well as infinite peaks at the data points, though this last issue is mitigated through a modified version of the estimate (Devroye & Krzyżak, 1999). Still, we find it interesting in the context of anomaly detection that weighting the edges may lead to introducing some dependence on the underlying density in the resulting scoring function. This contrasts with the unweighted case since at the limit the scoring function is agnostic to the density inside the support (the only remaining dependence is through the support).

Subsampling and scoring normalization Subsampling is a central component of the isolation forest method for anomaly detection. It is proposed in Liu et al. (2008) as a means of reducing the computational complexity of the method without drastically affecting the detection performance, and Liu et al. (2012) further advocate for using subsampling to mitigate the swamping and masking effects. In practice, a forest of isolation trees may be grown from subsamples of different sizes and the definition of normalized and interpretable anomaly scores requires proper scaling of the tree heights. This is the purpose of the normalization term $c(n)$ in the definition of the scoring function given in (5) that Liu et al. (2008), using an analogy between the structure of isolation trees and that of binary search trees, propose to take as $c(n) = 2\mathcal{H}_{n-1} - 2(n-1)/n = 2\mathcal{H}_n - 2$ as given in Preiss (1999) as the value of the average path length of unsuccesful searches in a binary search tree with n terminal nodes and which is understood as an estimation of the average heights in isolation trees. It is to be noted that, as such, $c(n)$ is actually defined as the average of the terminal nodes' heights further averaged according to the distribution of a binary search tree generated under the uniform random permutation model, where the binary search trees are grown by sequential insertion of a uniform random permutation of $\{1, \dots, n\}$, and the value of $c(n)$ is derived in Hibbard (1962). Although isolation trees and binary search trees share a binary recursive structure, their distribution differ in general. However, under a regular design in dimension 1, by Theorem 5 we have $\mathbb{E}[h_{\mathcal{T}}(x_1)] = \mathbb{E}[h_{\mathcal{T}}(x_n)] = \mathcal{H}_{n-1}$ and $\mathbb{E}[h_{\mathcal{T}}(x_i)] = \mathcal{H}_{i-1} + \mathcal{H}_{n-i}$ for $2 \leq i \leq n-1$, and this leads to $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[h_{\mathcal{T}}(x_i)] = \frac{2}{n}\sum_{i=1}^{n-1} \mathcal{H}_i = 2\mathcal{H}_{n-1} - 2(n-1)/n$, where we used the relation $\sum_{i=1}^{\ell} \mathcal{H}_i = (\ell+1)\mathcal{H}_\ell - \ell$. Therefore, the values of $c(n)$ and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[h_{\mathcal{T}}(x_i)]$ do agree in a one-dimensional regular design but this is not the case in an irregular design. That being said, $c(n)$ is equivalent to $2\log(n)$ as $n \to \infty$ and the same equivalence holds for $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[h_{\mathcal{T}}(x_i)]$ in either a random or fixed design by Corollary 9 and Theorem 10. Therefore $2\log(n)$ is the correct asymptotic scaling and Theorem 10 suggests an extra d factor in dimension d leading to a normalization by $2d\log(n)$ and to a range of asymptotic scoring values equal to $[0, 1/2]$ instead of $[0, 1]$.

Multi-dimensional considerations Our proof techniques are tied with the dimension being equal to one, as this induces a monotony property on the recursive partitioning that does not export well in dimension larger than one. Indeed, in proving Theorem 5 we make use of the fact that given n ordered points $x_1 < \cdots < x_n$, the isolation of an interior point x_i (with $2 \leq i \leq n-1$) is effective if and only if the recursive partition sequence contains a split between x_{i-1} and x_i and a split between x_i and x_{i+1}. This is also apparent in the bijection that we introduced between isolation tree restricted to the data and binary search trees used to store the indices of the intervals that contain the split points. But in dimension larger than one, even in a design with n points that project on each axis to n distinct values, the isolation of an interior point x does not necessarily require the occurence of a split between x and all its immediate neighbours. In fact, the isolation of a data point may become effective in a number of ways and so a multivariate analogue to Theorem 5 giving explicit expressions for the average heights in the case of a generic configuration of points seems difficult to obtain due to the combinatorial nature of the problem. However we conjecture that a convergence result along the line of Theorem 10 would hold in arbitrary dimension. For instance in the case of a random sample drawn from a distribution with a regular density f over $\mathbb{R}^d$ with compact support $\mathcal{S}$ and satisfying $\inf\{f(x) : x \in \mathcal{S}\} > 0$, we conjecture that $\mathbb{E}[h_{\mathcal{T}}(x)]/(d\log(n))$ would converge almost surely to 2 when $x \in \mathring{\mathcal{S}}$ and to a value strictly smaller than 2 when $x \notin \mathcal{S}^\circ$, although we anticipate that the function $x \mapsto \mathbb{E}[h_{\mathcal{T}}(x)]/(d\log(n))$ would exhibit a complex pattern, with potential discontinuities, even in the case of a smooth enough boundary $\partial\mathcal{S}$. One possible route may lie in considering a regular design in $\mathcal{S}$ that would serve as quantization points for the random sample and to derive perturbation bounds. We leave this interesting question as a perspective for future work.

8 Proofs

This section is organized as follows. Section 8.1 is devoted to the proofs of the non asymptotic results. This includes Proposition 2, Lemma 3, Theorem 5 and Proposition 6. The proofs for the asymptotic results in a random design (Theorem 7 and Theorem 8) are exposed in section 8.2, and those of the asymptotic results in a fixed design (Theorem 10 and Theorem 11) are presented in section 8.3.

8.1 Non asymptotic setting

8.1.1 Proof of Proposition 2

We recall first the bijections $\iota : \mathbb{T}_n \to \mathcal{B}_{2n-1}$ and $\tilde{\iota} : \mathcal{B}^S_{n-1} \to \mathcal{B}_{2n-1}$. We also recall the labelling $L^{\mathcal{B}}_{\mathcal{T}} : \mathcal{T}^\circ \to \{1, \dots, n-1\}$ defined for each $\mathcal{T} \in \mathcal{B}_{2n-1}$ by (7), and we point out that it is consistent both with the $\mathcal{D}_n$-restricted isolation trees in $\mathbb{T}_n$ and with the binary search trees in $\mathcal{B}^S_{n-1}$ in the following sense. For any $(\mathcal{T}, \pi_{\mathcal{T},n})$ in $\mathbb{T}_n$, at any internal node $v \in \mathcal{T}^\circ$ we have $L^{\mathcal{B}}_{\mathcal{T}}(v) = k$ if and only if there exists $\tau \in I_k$ that partitions the cell at v into the cells associated with its two children, meaning that $\pi_{\mathcal{T},n}(v0) = \pi_{\mathcal{T},n}(v) \cap (-\infty, \tau)$ and $\pi_{\mathcal{T},n}(v1) = \pi_{\mathcal{T},n}(v) \cap (\tau, \infty)$. And for a binary search tree $(\mathcal{T}, L_{\mathcal{T}}) \in \mathcal{B}^S_{n-1}$, we have $L^{\mathcal{B}}_{\mathcal{T}} = L_{\mathcal{T}}$, so that $L^{\mathcal{B}}_{\mathcal{T}}$ recovers the keys that are stored in the internal nodes of $\mathcal{T}$. We also note that during the generation of an isolation tree according to Algorithm 1 or a randomized binary search tree according to Algorithm 2, the construction of the left and right subtrees of a node v are conditionally independent given the path from the root to v.

For any integer k, we define the intervals $\mathcal{J}_0(k) = (-\infty, k]$ and $\mathcal{J}_1(k) = (k, +\infty)$. Given a subset $A \subset \mathbb{R}$ and $\eta \in \{0, 1\}$, we denote by A^η the set defined by $A^\eta = A$ if $\eta = 1$ and $A^\eta = \emptyset$ if $\eta = 0$.

Let $(\mathcal{T}^\star, \pi_{\mathcal{T}^\star,n}) \in \mathbb{T}_n$ be a $\mathcal{D}_n$-restricted isolation tree and let $(\widetilde{\mathcal{T}}^\star, L_{\widetilde{\mathcal{T}}^\star}) = (\tilde{\iota}^{-1} \circ \iota)\left((\mathcal{T}^\star, \pi_{\mathcal{T}^\star,n})\right)$ be its image by $\tilde{\iota}^{-1} \circ \iota$. Note that $\widetilde{\mathcal{T}}^\star = \mathcal{T}^\star$ and that $L_{\widetilde{\mathcal{T}}^\star} = L^{\mathcal{B}}_{\mathcal{T}^\star}$. We shall prove by recursion on the internal nodes that

$$\mu_n\left(\left[(\mathcal{T}^\star, \pi_{\mathcal{T}^\star,n})\right]\right) = \tilde{\mu}_n\left(\left[\left(\widetilde{\mathcal{T}}^\star, L_{\widetilde{\mathcal{T}}^\star}\right)\right]\right). \tag{26}$$

Let $(\mathcal{T}, \pi_{\mathcal{T},n}) \sim \mu_n$ be a generic random $\mathcal{D}_n$-restricted isolation tree taking values in $\mathbb{T}_n$ and let $(\widetilde{\mathcal{T}}, L_{\widetilde{\mathcal{T}}}) := (\widetilde{\mathcal{T}}, L_{\widetilde{\mathcal{T}}})(Z_1, \dots, Z_{n-1}) \sim \tilde{\mu}_n$ be a random binary search tree taking values in $\mathcal{B}^S_{n-1}$ and generated from independent random variables $Z_1, \dots, Z_{n-1}$ where Z_i is distributed according to $F_0^{w_i}$, for $1 \le i \le n-1$.

Let $i^\star_0 = L^{\mathcal{B}}_{\mathcal{T}^\star}(\emptyset)$. To initate the recursion, we need to prove that the event that the cell of the root node of $\mathcal{T}$ is partitioned at some point that belongs to $I_{i^\star_0}$ and the event that the key stored at the root of $\widetilde{\mathcal{T}}$ is $i^\star_0$ occur with the same probability, meaning that

$$\mathbb{P}\left(L^{\mathcal{B}}_{\mathcal{T}}(\emptyset) = i^\star_0\right) = \mathbb{P}\left(L_{\widetilde{\mathcal{T}}}(\emptyset) = i^\star_0\right). \tag{27}$$

On the one hand, the event that the first split point belongs to $I_{i^\star_0}$ occurs with probability $\frac{w_{i^\star_0}}{\sum_{\ell=1}^{n-1} w_\ell}$. On the other hand, by Lemma 3, the first key stored in $\widetilde{\mathcal{T}}$ is $i^\star_0$ if and only if $Z_{i^\star_0}$ is the largest among $Z_1, \dots, Z_{n-1}$, and this occurs with probability $\frac{w_{i^\star_0}}{\sum_{\ell=1}^{n-1} w_\ell}$ by Lemma 12. This proves (27).

Next, given $k \ge 1$, let $v^\star_k \in \mathcal{T}^{\star,\circ}$ be an internal node of $\mathcal{T}^\star$ and with height equal to k, so that $|v_k| = k$. Denote by $\emptyset = v^\star_0, v^\star_1, \dots, v^\star_k$ the (unique) shortest path from the root to $v^\star_k$ (notice that each $v^\star_\ell$, for each $0 \le \ell \le k$ is the ℓ-tuple composed of the first ℓ components of $v^\star_k$ since $v^\star_k$ is a k-tuple in the set of labels $\mathcal{U}$). Along this path in $(\mathcal{T}^\star, \pi_{\mathcal{T}^*,n})$, each cell $\pi_{\mathcal{T}^\star,n}(v^\star_\ell)$ is partitioned by a point that belongs to the interval at index $L^{\mathcal{B}}_{\mathcal{T}}(v^\star_\ell)$, for $0 \le \ell \le k$, and along this same path in $(\widetilde{\mathcal{T}}^\star, L_{\widetilde{\mathcal{T}}^\star})$, the key that is stored at $v^\star_\ell$ is $L^{\mathcal{B}}_{\mathcal{T}}(v^\star_\ell)$, for $0 \le \ell \le k$. Let $\Omega_k = [v^\star_k \in \mathcal{T}]$ be the event that $\mathcal{T}$ contains $v^\star_k$ and likewise, let $\widetilde{\Omega}_k = \left[v^\star_k \in \widetilde{\mathcal{T}}\right]$ be the event that $\widetilde{\mathcal{T}}$ contains $v^\star_k$.

On the event Ω_k, we have $L^{\mathcal{B}}_{\mathcal{T}}(v^\star_\ell) = L_{\mathcal{T}^\star}(v^\star_\ell)$ for all $0 \le \ell \le k$ and similarly on the event $\widetilde{\Omega}_k$, we have $L_{\widetilde{\mathcal{T}}}(v^\star_\ell) = L_{\mathcal{T}^\star}(v^\star_\ell)$ for all $0 \le \ell \le k$. Moreover, on the event Ω_k, for any $\eta \in \{0,1\}$, if $v^\star_k\eta \in \mathcal{T}^{\star,\circ}$ then $v^\star_k\eta \in \mathcal{T}^\circ$. Likewise, on the event $\widetilde{\Omega}_k$, for any $\eta \in \{0,1\}$, if $v^\star_k\eta \in \mathcal{T}^{\star,\circ}$ then $v^\star_k\eta \in \widetilde{\mathcal{T}}^\circ$. Therefore, the recursion will be established if we show that for any $\eta \in \{0,1\}$, whenever $v^\star_k\eta \in \mathcal{T}^{\star,\circ}$, we have

$$\mathbb{P}\left(L^{\mathcal{B}}_{\mathcal{T}}(v^\star_k\eta) = L^{\mathcal{B}}_{\mathcal{T}^\star}(v^\star_k\eta) \,|\, \Omega_k\right) = \mathbb{P}\left(L_{\widetilde{\mathcal{T}}}(v^\star_k\eta) = L^{\mathcal{B}}_{\mathcal{T}^\star}(v^\star_k\eta) \,|\, \widetilde{\Omega}_k\right), \tag{28}$$

meaning that the conditional probability that the cell $\pi_{\mathcal{T},n}(v^\star_k\eta)$ is partitioned by some point falling in the interval with index $L^{\mathcal{B}}_{\mathcal{T}^\star}(v^\star_k\eta)$ given Ω_k is equal to the conditional probability that the key $L^{\mathcal{B}}_{\mathcal{T}^\star}(v^\star_k\eta)$ is stored in the node $v^\star_k\eta$ of $\widetilde{\mathcal{T}}$ given $\widetilde{\Omega}_k$.

Now letting $i^\star_\ell = L_{\mathcal{T}^\star}(v^\star_\ell)$, for $0 \le \ell \le k$, on the event Ω_k, for any $\eta \in \{0,1\}$ such that $v^\star_k\eta \in \mathcal{T}^{\star,\circ}$, the cell $\pi_{\mathcal{T},n}(v^\star_k\eta)$ may be expressed explicitely in terms of the data points as $\pi_{\mathcal{T},n}(v^\star_k\eta) = \{x_\ell : \ell \in \mathcal{K}_\eta\}$, where

$$\mathcal{K}_\eta = \{1,\dots,n\} \cap \left(\bigcap_{j=1}^{k-1} \left(\mathcal{J}_0(i^\star_j)^{\mathbf{1}\{i^\star_{j+1}<i^\star_j\}} \cup \mathcal{J}_1(i^\star_j)^{\mathbf{1}\{i^\star_{j+1}>i^\star_j\}}\right)\right) \\ \cap \left(\mathcal{J}_0(i^\star_k)^{1-\eta} \cup \mathcal{J}_1(i^\star_k)^{\eta}\right).$$

Notice that $j^\star := L^{\mathcal{B}}_{\mathcal{T}^\star}(v^\star_k\eta)$ belongs to $\mathcal{K}_\eta \setminus (\max \mathcal{K}_\eta)$. Consequently, the conditional probability that $\pi_{\mathcal{T},n}(v^\star_k\eta)$ is partitioned by some point that belongs to the interval $I_{j^\star}$ given Ω_k is equal to $\frac{w_{j^\star}}{\sum_{\ell \in \mathcal{K}_\eta \setminus (\max \mathcal{K}_\eta)} w_\ell}$. Next, given $\widetilde{\Omega}_k$, the key stored at node $v^\star_k$ happens to be $j^\star$ if and only if $Z_{j^\star}$ is the largest among $\{Z_\ell : \ell \in \mathcal{K}_\eta\}$ by Lemma 3, and by Lemma 12, the conditional probability that this occurs given $\widetilde{\Omega}_k$ is equal to $\frac{w_{j^\star}}{\sum_{\ell \in \mathcal{K}_\eta \setminus (\max \mathcal{K}_\eta)} w_\ell}$. This proves (28) and the recursion is established. Then (27) and (28) yields (26) and the proof is complete.

8.1.2 Proof of Lemma 3

We consider that the Z_k's are all distinct since this holds with probability one. We expand the notation $\mathcal{T}^S(Z_1,\dots,Z_{n-1})$ into

$$\mathcal{T}^S(Z_1,\dots,Z_{n-1}) = (\mathcal{T}, L_{\mathcal{T}})(Z_1,\dots,Z_{n-1}),$$

and we refer to the internal nodes of $\mathcal{T}$ by the intervals $I_1,\dots,I_{n-1}$, meaning that node $v \in \mathcal{T}^\circ$ is referred to as I_i with $i = L_{\mathcal{T}}(v)$. Let $\mathcal{I} = \{\ell : i \wedge j \le \ell \le i \vee j\}$.

Suppose that Z_i is the largest among $\{Z_\ell : \ell \in \mathcal{I}\}$. If I_i is the root of $\mathcal{T}$, then obviously I_i is an ancestor of I_j. Else, denote by I_m the root of $\mathcal{T}$, and by $\mathcal{T}_0$ and $\mathcal{T}_1$ the (possibly empty) left and right subtrees of I_m in $\mathcal{T}$, respectively. By construction of $\mathcal{T}$, the internal nodes of $\mathcal{T}_0$ (respectively $\mathcal{T}_1$) are those intervals I_ℓ with $\ell < m$ (respectively $\ell > m$). Therefore necessarily either $m < i \wedge j$ or $m > i \vee j$, implying that the set of nodes $\{I_\ell : \ell \in \mathcal{I}\}$ lies entirely in $\mathcal{T}_0$ or in $\mathcal{T}_1$. We then proceed recursively on either $\mathcal{T}_0$ or $\mathcal{T}_1$ accordingly until I_i is found to be the root of the considered subtree, to conlude that I_i is an ancestor of I_j.

Conversely, suppose that I_i is an ancestor of I_j. If I_i is the root of $\mathcal{T}$, then Z_i is the largest of all the Z_ℓ's, and in particular among $\{Z_\ell : \ell \in \mathcal{I}\}$. Else, denote by I_m the root of $\mathcal{T}$, and by $\mathcal{T}_0$ and $\mathcal{T}_1$ the (possibly empty) left and right subtrees of I_m in $\mathcal{T}$, respectively, as above. Then necessarily either $m < i \wedge j$ or $m > i \vee j$, for otherwise $I_{i\wedge j}$ lies in $\mathcal{T}_0$ while $I_{i\vee j}$ lies in $\mathcal{T}_1$, implying that I_i is not an ancestor of I_j, hence a contradiction. Thus the set of nodes $\{I_\ell : \ell \in \mathcal{I}\}$ lies entirely in $\mathcal{T}_0$ or in $\mathcal{T}_1$, and by proceeding recursively on either $\mathcal{T}_0$ or $\mathcal{T}_1$ until I_i is found to be the root, we conlude that Z_i is the largest among $\{Z_\ell : \ell \in \mathcal{I}\}$.

8.1.3 Proof of Lemma 4

Recall the two sets $\mathcal{J}_1 = \{j : 1 \leq j \leq i-2\}$ and $\mathcal{J}_2 = \{j : i+1 \leq j \leq n-1\}$ of integers, and that we use the convention that a sum over an empty set is equal to 0. Using (10), we have

$$\begin{aligned}\max\{D_{i-1}, D_i\} &= 1 + \sum_{j\in\mathcal{J}_1} A_{j,i-1}A_{i,i-1} + A_{i,i-1} + \sum_{j\in\mathcal{J}_2} A_{j,i-1}A_{i,i-1} \\ &\quad + \sum_{j\in\mathcal{J}_1} A_{ji}A_{i-1,i} + A_{i-1,i} + \sum_{j\in\mathcal{J}_2} A_{ji}A_{i-1,i} \\ &= 2 + \sum_{j\in\mathcal{J}_1} \left(A_{j,i-1}A_{i,i-1} + A_{ji}A_{i-1,i}\right) \\ &\quad + \sum_{j\in\mathcal{J}_2} \left(A_{j,i-1}A_{i,i-1} + A_{ji}A_{i-1,i}\right),\end{aligned}$$

where we used the fact that $A_{i,i-1} + A_{i-1,i} = 1$.

Recall that $2 \leq i \leq n-1$ and note that $\mathcal{J}_1 = \emptyset$ when $i = 2$ and that $\mathcal{J}_2 = \emptyset$ when $i = n-1$. When $\mathcal{J}_1$ is not empty, for any $j \in \mathcal{J}_1$, we have that $A_{ji}A_{i-1,i} = 1$ if and only if $A_{j,i-1}A_{i-1,i} = 1$, meaning that I_j and I_i are common ancestors of I_{i-1} occurs if and only if I_j is an ancestor of I_{i-1} and I_{i-1} is an ancestor of I_i. This follows from the natural ordering of the intervals. To be sure, using the characterization of Lemma 3, we have

$$\begin{aligned}\left[A_{ji}A_{i-1,i} = 1\right] &= \left[Z_j \geq \max_{j\leq\ell\leq i} Z_\ell\right] \cap \left[Z_{i-1} \geq Z_i\right] \\ &= \left[Z_j \geq \max_{j\leq\ell\leq i-1} Z_\ell\right] \cap \left[Z_{i-1} \geq Z_i\right] \\ &= \left[A_{j,i-1}A_{i-1,i} = 1\right].\end{aligned}$$

Therefore, we have

$$\begin{aligned}\sum_{j\in\mathcal{J}_1} \left(A_{j,i-1}A_{i,i-1} + A_{ji}A_{i-1,i}\right) &= \sum_{j\in\mathcal{J}_1} A_{j,i-1}\left(A_{i,i-1} + A_{i-1,i}\right) \\ &= \sum_{j\in\mathcal{J}_1} A_{j,i-1}.\end{aligned}$$

Likewise, when $\mathcal{J}_2$ is not empty, for any $j \in \mathcal{J}_2$, $A_{j,i-1}A_{i,i-1} = 1$ if and only if $A_{ji}A_{i,i-1} = 1$, and this leads to $\sum_{j\in\mathcal{J}_2} \left(A_{j,i-1}A_{i,i-1} + A_{ji}A_{i-1,i}\right) = \sum_{j\in\mathcal{J}_2} A_{ji}$. Combining the above leads to the desired result.

8.1.4 Proof of Theorem 5

Recall first the operator $\Pi_n : (\mathcal{T}, \pi_\mathcal{T}) \mapsto (\mathcal{T}, \pi_{\mathcal{T},n})$ mapping an isolation tree $(\mathcal{T}, \pi_\mathcal{T})$ in $\mathbb{T}$ to its restriction $(\mathcal{T}, \pi_{\mathcal{T},n})$ to $\mathcal{D}_n$, and that $(h_\mathcal{T}(x_1), \dots, h_\mathcal{T}(x_n)) = \left(h_{\mathcal{T},n}(x_1), \dots, h_{\mathcal{T},n}(x_n)\right)$ for any $(\mathcal{T}, \pi_\mathcal{T}) \in \mathbb{T}$ such that $(\mathcal{T}, \pi_{\mathcal{T},n}) = \Pi_n\left((\mathcal{T}, \pi_\mathcal{T})\right)$. We use this in conjunction with Proposition 2, equations (8) and (9) and Lemma 4 to relate the expectations of $h_{\mathcal{T},n}(x_i)$, $1 \leq i \leq n$, with the expectations of the ancestor variables and some of their products, to deduce that

$$h_\mathcal{T}(x_i) \stackrel{\mathcal{L}}{=} \begin{cases} 1 + \sum_{j=1}^{n} A_{j1} & \text{if } i = 1, \\ 2 + \sum_{j=1}^{i-2} A_{j,i-1} + \sum_{j=i+1}^{n-1} A_{ji} & \text{if } 2 \leq i \leq n-1, \\ 1 + \sum_{j=1}^{n-1} A_{j,n-1} & \text{if } i = n, \end{cases} \tag{29}$$

where we use the convention that a sum over an empty set of indices is equal to 0. We recall that the statements involving ancestor variables relate to a generic random binary search tree $\mathcal{T}^S := \mathcal{T}^S(Z_1, \dots, Z_{n-1})$ where $Z_1, \dots, Z_{n-1}$ are independent random variables with $Z_i \sim F_0^{w_i}$, for $1 \leq i \leq n-1$, so that $\mathcal{T}^S$ is distributed according to $\tilde{\mu}_n$.

Using the characterization of the ancestor variables given in Lemma 3, we note that the two sums in the right hand side of (29) for the case where $2 \leq i \leq n-1$ are independent. In addition, Lemma 3 and Lemma 13 combined imply that the set of ancestor variables involved in each sum in (29) are independent as well. To complete the proof, there remains to evaluate the expectations of the ancestor variables. Applying Lemma 12, we directly obtain that

$$\mathbb{E}[A_{ji}] = \mathbb{P}\left(Z_i \geq \max_{i \wedge j \leq \ell \leq i \vee j} Z_\ell\right) = \frac{w_j}{\sum_{i \wedge j \leq \ell \leq i \vee j} w_\ell}.$$

Using this together with (29) leads to the expressions given in (11) and (12).

8.1.5 Proof of Proposition 6

Let $x = (x^{(1)}, \ldots, x^{(d)}) \in \mathrm{Conv}(\mathcal{D}_n)$ be a point in the convex hull of $\mathcal{D}_n$. Note first that $\mathrm{Conv}(\mathcal{D}_n)$ is an orthotope (meaning a hyperrectangle) of dimension d, and that x is included in the orthotope with vertices $\{x_{\mathbf{i}} : \mathbf{i} \in \mathbf{I}(x)\}$. We denote by $w^j = x_{j,i_j(x)+1} - x_{j,i_j(x)}$, for $j \in \{1, \ldots, d\}$ the side lengths of this orthotope, where we recall that $i_j(x) = 1 + \lfloor (n-1)(x^{(j)} - x_{j,1})/(x_{j,n} - x_{j,1}) \rfloor$, for $j \in \{1, \ldots, d\}$. For $\mathbf{i} \in \mathbf{I}(x)$, we let $\delta_j(\mathbf{i}) = i_j - i_j(x)$ and we note that $\delta_j(\mathbf{i}) \in \{0, 1\}$.

Let $V_{\mathbf{I}(x)}(\mathcal{T}) = \{v \in \partial\mathcal{T} : \pi_{\mathcal{T}}(v) \cap \{x_{\mathbf{i}} : \mathbf{i} \in \mathbf{I}(x)\} \neq \emptyset\}$ be the set of leaves in $\mathcal{T}$ the attached cells of whose contain the points $x_{\mathbf{i}}$, for $\mathbf{i} \in \mathbf{I}(x)$ (each of those cells contain exactly one point). Notice that x necessarily belongs to one of the cells $\pi_{\mathcal{T}}(v)$ for $v \in V_{\mathbf{I}(x)}$, so that $h_{\mathcal{T}}(x)$ is equal to $h_{\mathcal{T}}(x_{\mathbf{i}})$ for some $\mathbf{i} \in \mathbf{I}(x)$, which depends on $\mathcal{T}$. To compute the expectation of $h_{\mathcal{T}}(x)$, we first condition on isolation trees restricted to $\mathcal{D}_n$, which prescribes the tree structure leaving free only the random draws of the split values. Indeed, recall that an isolation tree and its restriction to $\mathcal{D}_n$ carry the same tree structure, meaning that for any $(\mathcal{T}_0, \pi_{\mathcal{T}_0}) \in \mathbb{T}$ and $(\mathcal{T}'_0, \pi_{\mathcal{T}'_0,n}) \in \mathbb{T}_n$ such that $(\mathcal{T}'_0, \pi_{\mathcal{T}'_0,n}) = \Pi_n\left((\mathcal{T}_0, \pi_{\mathcal{T}_0})\right)$, we have $\mathcal{T}'_0 = \mathcal{T}_0$.

We have

$$\mathbb{E}[h_{\mathcal{T}}(x)] = \sum_{(\mathcal{T}', \pi_{\mathcal{T}',n}) \in \mathbb{T}_n} \mathbb{E}\left[h_{\mathcal{T}}(x) \middle| \left[\Pi_n\left((\mathcal{T}, \pi_{\mathcal{T}})\right) = \left(\mathcal{T}', \pi_{\mathcal{T}',n}\right)\right]\right] \times \mathbb{P}\left(\Pi_n\left((\mathcal{T}, \pi_{\mathcal{T}})\right) = \left(\mathcal{T}', \pi_{\mathcal{T}',n}\right)\right). \tag{30}$$

Let $(\mathcal{T}', \pi_{\mathcal{T}',n})$ be a $\mathcal{D}_n$-restricted isolation tree in $\mathbb{T}_n$ and let $V_{\mathbf{I}(x),n}(\mathcal{T}') = \{v \in \partial\mathcal{T}' : \pi_{\mathcal{T}',n}(v) \cap \{x_{\mathbf{i}} : \mathbf{i} \in \mathbf{I}(x)\} \neq \emptyset\}$. For any interior node $v \in \mathcal{T}^\circ$ of $\mathcal{T}$, denote by $j(v) \in \{1, \ldots, d\}$ and $\tau(v) \in \mathbb{R}$ respectively the split component and split value associated with v. Likewise, for any interior node $v \in \mathcal{T}'^\circ$, let $j'(v) \in \{1, \ldots, d\}$ and $i'(v) \in \{1, \ldots, n-1\}$ be such that $\pi_{\mathcal{T}',n}(v)$ is split between $x_{j'(v),i'(v)}$ and $x_{j'(v),i'(v)+1}$.

Consider the event $\Omega' = [\Pi_n\left((\mathcal{T}, \pi_{\mathcal{T}})\right) = (\mathcal{T}', \pi_{\mathcal{T}',n})]$. On Ω', we have $\mathcal{T} = \mathcal{T}'$, as well as $h_{\mathcal{T}}(x_{\mathbf{i}}) = h_{\mathcal{T}',n}(x_{\mathbf{i}})$ for any $\mathbf{i} \in \{1, \ldots, n\}^d$, and so in particular for any $\mathbf{i} \in \mathbf{I}(x)$. On Ω' it also holds that $V_{\mathbf{I}(x)}(\mathcal{T}) = V_{\mathbf{I}(x),n}(\mathcal{T}')$, that $j(v) = j'(v)$ and that $x_{j'(v),i'(v)} \leq \tau(v) \leq x_{j'(v),i'(v)+1}$, for any $v \in \mathcal{T}^\circ = \mathcal{T}'^\circ$.

Let $v^\star$ be the least common ancestor to all the nodes in $V_{\mathbf{I}(x),n}(\mathcal{T}')$, meaning the node with largest height which is an ancestor in $\mathcal{T}'$ of every node in $V_{\mathbf{I}(x),n}(\mathcal{T}')$. Notice that $v^\star$ is the node at which the points $\{x_{\mathbf{i}} : \mathbf{i} \in \mathbf{I}(x)\}$ part ways, so to speak, meaning some of these points belong to $\pi_{\mathcal{T}',n}(v^\star 0)$ while the remaining points belong to $\pi_{\mathcal{T}',n}(v^\star 1)$, where we recall that $v^\star 0$ and $v^\star 1$ denote the left and right children of $v^\star$. Therefore we have $i'(v^\star) = i_{j'(v^\star)}(x)$.

Then, letting $j^\star = j'(v^\star)$ and $i^\star = i_{j'(v^\star)}(x)$ to ease notation, we obtain that the conditional probability that $\pi_{\mathcal{T}}(v^\star 0)$ contains x given Ω' is equal to $1 - (x^{(j^\star)} - x_{j\star,iv^\star})/w^{j^\star})$ and similarly, that the conditional probability that $\pi_{\mathcal{T}}(v^\star 1)$ contains x is equal to $(x^{(j^\star)} - x_{j^\star,i^\star})/w^{j^\star}$, due to the fact that the conditional distribution of $\tau(v^\star)$ given Ω' is a uniform distribution over

$[x_{j^\star,i^\star}, x_{j^\star,i^\star+1}]$. By proceeding recursively on each subtree of $v^\star$ as we just did, we deduce that

$$\mathbb{E}[h_{\mathcal{T}}(x)|\Omega'] = \sum_{\mathbf{i}\in\mathbf{I}(x)} \alpha_{\mathbf{i}} h_{\mathcal{T}',n}(x_{\mathbf{i}}), \tag{31}$$

where $\alpha_{\mathbf{i}}$ is given by (14) and where we used the fact that the split values $\tau(v)$ for all $v \in \mathcal{T}^\circ$ are conditionally independent given Ω'. Importantly, the expression of $\alpha_{\mathbf{i}}$ in (14) does not depend on $(\mathcal{T}', \pi_{\mathcal{T}',n})$ but only on x. By reporting (31) in (30), we obtain that

$$\begin{aligned}\mathbb{E}[h_{\mathcal{T}}(x)] &= \sum_{\mathbf{i}\in\mathbf{I}(\mathbf{x})} \alpha_{\mathbf{i}} \sum_{(\mathcal{T}',\pi_{\mathcal{T}',n})\in\mathbb{T}_n} h_{\mathcal{T}',n}(x_{\mathbf{i}}) \mathbb{P}\left(\Pi_n\left((\mathcal{T},\pi_{\mathcal{T}})\right) = \left(\mathcal{T}',\pi_{\mathcal{T}',n}\right)\right) \\ &= \sum_{\mathbf{i}\in\mathbf{I}(\mathbf{x})} \alpha_{\mathbf{i}} \mathbb{E}[h_{\mathcal{T}}(x_{\mathbf{i}})],\end{aligned}$$

and this proves (13).

We now prove the second statement. Denote by $\Psi : \mathbb{R}^d \to \mathrm{Conv}(\mathcal{D}_n)$ the projection operator onto the convex hull of $\mathcal{D}_n$. During the construction of any isolation tree $(\mathcal{T}, \pi_{\mathcal{T}})$ according to Algorithm 1, the cells are partitioned based on a hyperplane orthogonal to one of the coordinate axis. This implies that for any leaf $v \in \partial\mathcal{T}$, the cell $\pi_{\mathcal{T}}(v)$ extends outside $\mathrm{Conv}(\mathcal{D}_n)$ in such a way that $\pi_{\mathcal{T}}(v) \supset \left\{x \in \mathbb{R}^d : \Psi(x) \in \pi_{\mathcal{T}}(v) \cap \mathcal{D}_n\right\}$. Consequently, since the points in $\mathcal{D}_n$ are arranged as a grid parallel to the coordinate axes, we have $\mathbb{E}[h_{\mathcal{T}}(x)] = \mathbb{E}\left[h_{\mathcal{T}}\left(\Psi(x)\right)\right]$ for any $x \in \mathbb{R}^d$.

8.2 Asymptotics in a random design

In this section we prove Theorem 7 and Theorem 8. First of all, we recall that $G = F^{-1}$ denotes the quantile function defined by $G(p) = \inf\{x : F(x) \geq p\}$, for $0 \leq p \leq 1$. Also, for any $p \in (0,1)$ at which G' and G'' exist, their expressions are given by $G'(p) = f(G(p))^{-1}$ and $G''(p) = -f'(G(p))/f(G(p))^3$ respectively.

8.2.1 Proof of Theorem 7, statement (i)

For ease of notation, we denote $i_n(p)$ by i, the dependence on n and on p being understood. Suppose that n is large enough that $2 \leq i \leq n-1$. Then from (19), we have

$$\bar{H}_i = \sum_{j=1}^{i-1} \frac{X_{(j+1)} - X_{(j)}}{X_{(i)} - X_{(j)}} + \sum_{j=i+1}^{n} \frac{X_{(j)} - X_{(j-1)}}{X_{(j)} - X_{(i)}} =: A_i + B_i. \tag{32}$$

We prove that both $\frac{1}{\log(n)} A_i$ and $\frac{1}{\log(n)} B_i$ converge to 1 almost surely as $n \to \infty$. Since the two series A_i and B_i have analogous expressions, we only prove the convergence for the right series B_i.

Let $r := r_n = \lfloor n/\log(n) \rfloor$ and suppose that n is large enough that $3 \leq r \leq n-i$. Then we decompose B_i into

$$B_i = \sum_{j=i+1}^{i+r} \frac{X_{(j)} - X_{(j-1)}}{X_{(j)} - X_{(i)}} + \sum_{j=i+r+1}^{n} \frac{X_{(j)} - X_{(j-1)}}{X_{(j)} - X_{(i)}} =: B_{i,1} + B_{i,2}, \tag{33}$$

and we prove that

$$\frac{1}{\log(n)} B_{i,1} \to 1 \qquad \text{almost surely as } n \to \infty, \tag{34}$$

and that

$$\frac{1}{\log(n)} B_{i,2} \to 0 \qquad \text{almost surely as } n \to \infty. \tag{35}$$

To this aim, we use the connection between the order statistics of $\{X_1, \dots, X_n\}$ and that of a uniform sample. Let $U_j = F(X_j)$, for $j = 1, \dots, n$. Then the random variables $U_1, \dots, U_n$ are independent and identically distributed according to a uniform distribution over $[0,1]$, and for any $1 \le j \le n$, we have $U_{(j)} = F(X_{(j)})$, where $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$ denote the order statistics of the sample $U_1, \dots, U_n$.

Proof of (34): convergence of $\frac{1}{\log(n)} B_{i,1}$. We have

$$B_{i,1} = \sum_{j=i+1}^{i+r} \frac{G(U_{(j)}) - G(U_{(j-1)})}{G(U_{(j)}) - G(U_{(i)})} = 1 + \sum_{j=i+2}^{i+r} \frac{G(U_{(j)}) - G(U_{(j-1)})}{G(U_{(j)}) - G(U_{(i)})}.$$

By continuity of f in a neighborhood of x_p, there exists $\delta > 0$ such that the application $y \mapsto f(G(y))$ is continuous and bounded away from 0 over the closed interval $[p-\delta, p+\delta]$. Hence G' and G'' are well defined over $[p-\delta, p+\delta]$. We define the following constants:

$$\begin{aligned} \kappa_1 &= \inf\{f(G(y)) \,:\, p-\delta \le y \le p+\delta\}, \\ \kappa_2 &= \sup\{f(G(y)) \,:\, p-\delta \le y \le p+\delta\}, \\ \kappa_3 &= \sup\{|f'(G(y))| \,:\, p-\delta \le y \le p+\delta\}, \end{aligned} \tag{36}$$

and we note that $\kappa_1 > 0$ and $\kappa_2 < \infty$. Then, for any $p-\delta \le y \le p+\delta$, we have

$$|G'(y)| \ge 1/\kappa_2 \quad \text{and} \quad |G''(y)| \le \kappa_3/\kappa_1^3.$$

Let $\mathcal{E}_{n,1}$ be the event defined by

$$\mathcal{E}_{n,1} = \left[U_{(i)} \ge p-\delta\right] \cap \left[U_{(i+r)} \le p+\delta\right].$$

On the event $\mathcal{E}_{n,1}$, we have $p-\delta \le U_{(j)} \le p+\delta$ for any $i \le j \le i+r$, and using Taylor expansions at $U_{(j-1)}$ and $U_{(i)}$ for each $i+2 \le j \le i+r$, we obtain that

$$\begin{aligned} G(U_{(j)}) &= G(U_{(j-1)}) + G'(U_{(j-1)})(U_{(j)} - U_{(j-1)}) + R_j, \\ G(U_{(j)}) &= G(U_{(i)}) + G'(U_{(i)})(U_{(j)} - U_{(i)}) + \tilde{R}_j, \end{aligned}$$

where the remainder terms R_j and $\tilde{R}_j$ are expressed respectively as

$$R_j = \frac{1}{2} G''(\xi_j)(U_{(j)} - U_{(j-1)})^2 \quad \text{and} \quad \tilde{R}_j = \frac{1}{2} G''(\tilde{\xi}_j)(U_{(j)} - U_{(i)})^2,$$

for some random variables ξ_j and $\tilde{\xi}_j$ taking values in $[p-\delta, p+\delta]$. Therefore, on the event $\mathcal{E}_{n,1}$, we have

$$B_{i,1} = 1 + \sum_{j=i+2}^{i+r} \frac{G'(U_{(j-1)})(U_{(j)} - U_{(j-1)})}{G'(U_{(i)})(U_{(j)} - U_{(i)}) + \tilde{R}_j} + \sum_{j=i+2}^{i+r} \frac{R_j}{G(U_{(j)}) - G(U_{(i)})}. \tag{37}$$

For each integer j with $i+2 \le j \le i+r$, we have

$$\frac{G'(U_{(j-1)})(U_{(j)} - U_{(j-1)})}{G'(U_{(i)})(U_{(j)} - U_{(i)}) + \tilde{R}_j} = \frac{G'(U_{(j-1)})}{G'(U_{(i)})} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} \frac{1}{1 + \frac{\tilde{R}_j}{G'(U_{(i)})(U_{(j)} - U_{(i)})}}, \tag{38}$$

and

$$\left| \frac{\tilde{R}_j}{G'(U_{(i)})(U_{(j)} - U_{(i)})} \right| \le \frac{1}{2} \frac{\left|G''(\tilde{\xi}_j)\right| (U_{(j)} - U_{(i)})^2}{G'(U_{(i)})(U_{(j)} - U_{(i)})} \le \frac{\kappa_2 \kappa_3}{2\kappa_1^3} (U_{(j)} - U_{(i)}),$$

where we used the bounds in (36). Let $\mathcal{E}_{n,2}$ be the event defined by

$$\mathcal{E}_{n,2} = \left[U_{(i+r)} - U_{(i)} \leq \frac{\kappa_1^3}{\kappa_2(\kappa_3 \vee 1)} \right]. \tag{39}$$

(We use the maximum $\kappa_3 \vee 1$ in (39) since κ_3 may be equal to 0 if f' vanishes in a neighborhood of x_p). Then, on the event $\mathcal{E}_{n,2}$, we have $\frac{\kappa_2\kappa_3}{2\kappa_1^3}(U_{(j)} - U_{(i)}) \leq \frac{1}{2}$, and using the inequality $|1/(1+x) - 1| \leq 2|x|$ over $[-1/2, 1/2]$, this yields

$$\left| \frac{1}{1 + \frac{\tilde{R}_j}{G'(U_{(i)})(U_{(j)} - U_{(i)})}} - 1 \right| \leq \frac{\kappa_2\kappa_3}{\kappa_1^3}(U_{(j)} - U_{(i)}). \tag{40}$$

By combining (40) and (38), and by summing over j, we obtain that on the event $\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}$,

$$\begin{aligned}
&\left| \sum_{j=i+2}^{i+r} \frac{G'(U_{(j-1)})(U_{(j)} - U_{(j-1)})}{G'(U_{(i)})(U_{(j)} - U_{(i)}) + \tilde{R}_j} - \sum_{j=i+1}^{i+r} \frac{G'(U_{(j-1)})}{G'(U_{(i)})} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} \right| \\
&\quad \leq \sum_{j=i+2}^{i+r} \frac{G'(U_{(j-1)})}{G'(U_{(i)})} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} \left| \frac{1}{1 + \frac{\tilde{R}_j}{G'(U_{(i)})(U_{(j)} - U_{(i)})}} - 1 \right| \\
&\quad \leq \frac{\kappa_2}{\kappa_1} \sum_{j=i+2}^{i+r} \frac{\kappa_2\kappa_3}{\kappa_1^3}(U_{(j)} - U_{(j-1)}) \\
&\quad \leq \frac{\kappa_2^2\kappa_3}{\kappa_1^4}(U_{(i+r)} - U_{(i)}).
\end{aligned} \tag{41}$$

For each $i+2 \leq j \leq i+r$, on the event $\mathcal{E}_{n,1}$, a Taylor expansion of G' at $U_{(i)}$ yields

$$G'(U_{(j-1)}) - G'(U_{(i)}) = G''(\eta_{j-1})(U_{(j-1)} - U_{(i)}),$$

for some random variables η_{j-1} taking values in $[p-\delta, p+\delta]$, so that on the event $\mathcal{E}_{n,1}$, we have

$$\begin{aligned}
&\left| \sum_{j=i+2}^{i+r} \frac{G'(U_{(j-1)})}{G'(U_{(i)})} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} - \sum_{j=i+2}^{i+r} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} \right| \\
&\quad \leq \sum_{j=i+2}^{i+r} \left| \frac{G'(U_{(j-1)}) - G'(U_{(i)})}{G'(U_{(i)})} \right| \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} \\
&\quad \leq \frac{\kappa_2\kappa_3}{\kappa_1^3} \sum_{j=i+2}^{i+r} (U_{(j-1)} - U_{(i)}) \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} \\
&\quad \leq \frac{\kappa_2\kappa_3}{\kappa_1^3} \sum_{j=i+2}^{i+r} (U_{(j)} - U_{(j-1)}) \\
&\quad \leq \frac{\kappa_2\kappa_3}{\kappa_1^3}(U_{(i+r)} - U_{(i)}),
\end{aligned} \tag{42}$$

where we used the facts that $U_{(j)} - U_{(i)} \geq 0$ and that $\frac{U_{(j-1)} - U_{(i)}}{U_{(j)} - U_{(i)}} \leq 1$. By combining (41) and (42), it follows that on $\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}$,

$$\left| \sum_{j=i+2}^{i+r} \frac{G'(U_{(j-1)})(U_{(j)} - U_{(j-1)})}{G'(U_{(i)})(U_{(j)} - U_{(i)}) + \tilde{R}_j} - \sum_{j=i+2}^{i+r} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} \right| \leq c_1 \left(U_{(i+r)} - U_{(i)} \right), \tag{43}$$

for some constant $c_1 > 0$ depending only on f and which can be taken as $c_1 = \frac{\kappa_2\kappa_3}{\kappa_1^3}\left(1 + \frac{\kappa_2}{\kappa_1}\right)$.

Now we bound the second sum in the expression of $B_{i,1}$ given in (37). On the event $\mathcal{E}_{n,1}$, we have

$$\sum_{j=i+2}^{i+r} \frac{|R_j|}{G(U_{(j)}) - G(U_{(i)})} = \frac{1}{2} \sum_{j=i+2}^{i+r} \frac{|G''(\xi_j)|(U_{(j)} - U_{(j-1)})^2}{G(U_{(j)}) - G(U_{(i)})} \leq \frac{\kappa_3}{2\kappa_1^3} \sum_{j=i+2}^{i+r} \frac{(U_{(j)} - U_{(j-1)})^2}{G(U_{(j)}) - G(U_{(j-1)})},$$

where we used the bounds (36) together with the fact that $G(U_{(j)}) - G(U_{(i)}) \geq G(U_{(j)}) - G(U_{(j-1)})$. On the event $\mathcal{E}_{n,1}$, for each $i+2 \leq j \leq i+r$, the ratio $\frac{U_{(j)} - U_{(j-1)}}{G(U_{(j)}) - G(U_{(j-1)})}$, which is positive, is bounded from above by κ_2 by the mean value theorem. Therefore, on $\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}$, we have

$$\sum_{j=i+2}^{i+r} \frac{|R_j|}{G(U_{(j)}) - G(U_{(i)})} \leq \frac{\kappa_2 \kappa_3}{2\kappa_1^3} \sum_{j=i+2}^{i+r} (U_{(j)} - U_{(j-1)}) \leq \frac{\kappa_2 \kappa_3}{2\kappa_1^3} (U_{(i+r)} - U_{(i)}). \tag{44}$$

Thus, by combining (43) and (44), we have shown that there exists constants $c_2 > 0$ and $c_3 > 0$ depending only on f such that, on $\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}$,

$$\left| B_{i,1} - \sum_{j=i+1}^{i+r} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} \right| \leq c_2 \left(U_{(i+r)} - U_{(i)} \right) \leq c_3, \tag{45}$$

where we used the fact that $U_{(i+r)} - U_{(i)} \leq 2\frac{\kappa_1^3}{\kappa_2(\kappa_3 \vee 1)} \wedge 2\delta$ on $\mathcal{E}_{n,1} \cap \mathcal{E}_{n,2}$, and where c_2 and c_3 can be taken respectively as $c_2 = c_1 + \frac{\kappa_2 \kappa_3}{2\kappa_1^3}$ and as $c_3 = 2\delta c_2$.

For any real number $t > 0$, and for all n large enough that $3 \leq r \leq n - i$ (which we assumed in (33)), we have

$$\mathbb{P}\left(\left| \frac{B_{i,1}}{\log(n)} - 1 \right| > t \right) \leq \mathbb{P}\left(\left[\left| \frac{B_{i,1}}{\log(n)} - 1 \right| > t \right] \cap \mathcal{E}_{n,1} \cap \mathcal{E}_{n,2} \right) + \mathbb{P}\left(\mathcal{E}_{n,1}^c \right) + \mathbb{P}\left(\mathcal{E}_{n,2}^c \right). \tag{46}$$

Using (45), we bound the first term in (46) as

$$\mathbb{P}\left(\left[\left| \frac{B_{i,1}}{\log(n)} - 1 \right| > t \right] \cap \mathcal{E}_{n,1} \cap \mathcal{E}_{n,2} \right) \leq \mathbb{P}\left(\left| \frac{1}{\log(n)} \sum_{j=i+1}^{i+r} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} - 1 \right| > \frac{t}{2} \right), \tag{47}$$

for all n large enough that $c_3/\log(n) < t/2$. With our choice of $r = \lfloor n/\log(n) \rfloor$, we have $\log(r) < \log(n)$, so that

$$\left| \frac{1}{\log(n)} \sum_{j=i+1}^{i+r} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} - 1 \right| \leq \left| \frac{1}{\log(r)} \sum_{j=i+1}^{i+r} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} - 1 \right| + \left| 1 - \frac{\log(n)}{\log(r)} \right|.$$

Hence, since $\log(r)/\log(n) \to 1$ as $n \to \infty$, we obtain that for any $t > 0$ and for all n large enough that $|1 - \frac{\log(n)}{\log(r)}| < t/4$,

$$\mathbb{P}\left(\left| \frac{1}{\log(n)} \sum_{j=i+1}^{i+r} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} - 1 \right| > \frac{t}{2} \right) \leq \mathbb{P}\left(\left| \frac{1}{\log(r)} \sum_{j=i+1}^{i+r} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}} - 1 \right| > \frac{t}{4} \right). \tag{48}$$

To bound the probability above, we use the representation of uniform order statistics with exponential random variables (see for instance Ahsanullah et al., 2013, Chapter 4). Let $(\nu_j)_{(j\geq 1)}$ be an IID sequence of standard exponential random variables. Then

$$\left(U_{(1)},\dots,U_{(n)}\right) \stackrel{\mathcal{L}}{=} \left(\frac{\nu_1}{\nu_1+\cdots+\nu_{n+1}},\dots,\frac{\nu_1+\cdots+\nu_n}{\nu_1+\cdots+\nu_{n+1}}\right), \tag{49}$$

so that

$$\sum_{j=i+1}^{i+r}\frac{U_{(j)}-U_{(j-1)}}{U_{(j)}-U_{(i)}} \stackrel{\mathcal{L}}{=} \sum_{j=i+1}^{i+r}\frac{\nu_j}{\nu_{i+1}+\cdots+\nu_j} \stackrel{\mathcal{L}}{=} \sum_{j=1}^{r}\frac{\nu_j}{\nu_1+\cdots+\nu_j}, \tag{50}$$

where $\stackrel{\mathcal{L}}{=}$ means that the terms on each side of the equal sign have the same distribution. Using this, together with Lemma 14, we obtain that, for any $t>0$ and for all n large enough,

$$\begin{aligned}\mathbb{P}&\left(\left|\frac{1}{\log(r)}\sum_{j=i+1}^{i+r}\frac{U_{(j)}-U_{(j-1)}}{U_{(j)}-U_{(i)}}-1\right|>\frac{t}{4}\right)\\ &\leq 2\exp\left(-\frac{t}{16}\log(r)\log\left(1+\frac{3t\log(r)}{4\pi^2}\right)\right).\end{aligned} \tag{51}$$

When n is large enough that $\log(r)>\log(n)/2$, the right-hand side in (51) is bounded by $2\exp\left(-\frac{t}{32}\log(n)\log\left(1+\frac{3t\log(n)}{8\pi^2}\right)\right)$, which in turn is bounded by $\frac{2}{n^2}$ when n is large enough that $\frac{t}{32}\log\left(1+\frac{3t\log(n)}{8\pi^2}\right)>2$. Using this bound together with (48) and (51), we deduce that for any $t>0$,

$$\sum_{n\geq 1}\mathbb{P}\left(\left[\left|\frac{B_{i,1}}{\log(n)}-1\right|>t\right]\cap\mathcal{E}_{n,1}\cap\mathcal{E}_{n,2}\right)<\infty. \tag{52}$$

Now we prove that both series $\sum_{n\geq 1}\mathbb{P}(\mathcal{E}_{n,1}^c)$ and $\sum_{n\geq 1}\mathbb{P}(\mathcal{E}_{n,2}^c)$ are convergent. To this aim, we introduce the uniform empirical quantile process $\mathbb{F}_n^{-1}(p)=\inf\{y:\mathbb{F}_n(y)\geq p\}$, for $0\leq p\leq 1$, where $\mathbb{F}_n(t)=\frac{1}{n}\sum_{j=1}^n\mathbf{1}\{U_j\leq t\}$ for $0\leq t\leq 1$. We start with the event $\mathcal{E}_{n,1}$. We have $U_{(i)}=\mathbb{F}_n^{-1}\left(\frac{i}{n}\right)=\mathbb{F}_n^{-1}\left(\frac{i}{n}\right)-\frac{i}{n}+\frac{i}{n}-p+p\geq -\|\mathbb{F}_n^{-1}-I\|_\infty-\frac{\delta}{2}+p$, where the inequality holds for all n large enough that $|\frac{i}{n}-p|\leq\frac{\delta}{2}$. Hence, for all n large enough

$$\mathbb{P}\left(U_{(i)}\geq p-\delta\right)\geq\mathbb{P}\left(\|\mathbb{F}_n^{-1}-1\|_\infty\leq\frac{\delta}{2}\right).$$

Likewise, $U_{(i+r)}=\mathbb{F}_n^{-1}\left(\frac{i+r}{n}\right)\leq\|\mathbb{F}_n^{-1}-I\|_\infty+\frac{\delta}{2}+p$, where the inequality holds for all n large enough that $|\frac{i+r}{n}-p|\leq\frac{\delta}{2}$, and this yields

$$\mathbb{P}\left(U_{(i+r)}\leq p+\delta\right)\geq\mathbb{P}\left(\|\mathbb{F}_n^{-1}-1\|_\infty\leq\frac{\delta}{2}\right),$$

for all n large enough. Hence, for all n large enough, we have

$$\mathbb{P}\left(\mathcal{E}_{n,1}^c\right)\leq 2\mathbb{P}\left(\|\mathbb{F}_n^{-1}-1\|_\infty\geq\frac{\delta}{2}\right)\leq 4\exp\left(-\frac{n\delta^2}{2}\right),$$

where we used Proposition 15 to bound the supremum of the empirical quantile process. Consequently,

$$\sum_{n\geq 1}\mathbb{P}\left(\mathcal{E}_{n,1}^c\right)<\infty. \tag{53}$$

Now we turn to the sequence of events $\mathcal{E}_{n,2}$. Set $c=\frac{\kappa_1^3}{\kappa_2(\kappa_3\vee 1)}$, and we recall that $\mathcal{E}_{n,2}$ is the event that $U_{(i+r)}-U_{(i)}\leq c$. We have $U_{(i+r)}-U_{(i)}=\mathbb{F}_n^{-1}\left(\frac{i+r}{n}\right)-\mathbb{F}_n^{-1}\left(\frac{i}{n}\right)\leq 2\|\mathbb{F}_n^{-1}-I\|_\infty+\frac{r}{n}$, and using Proposition 15 as above, we obtain that for all n large enough that $\frac{r}{n}\leq\frac{c}{3}$,

$$\mathbb{P}\left(\mathcal{E}_{n,2}^c\right)\leq\mathbb{P}\left(\|\mathbb{F}_n^{-1}-1\|_\infty\geq\frac{c}{3}\right)\leq 2\exp\left(-\frac{2nc^2}{9}\right),$$

which yields that

$$\sum_{n\geq 1} \mathbb{P}\left(\mathcal{E}_{n,2}^c\right) < \infty. \tag{54}$$

Combining (52), (53), and (54) with (46), we deduce that, for any $t > 0$,

$$\sum_{n\geq 1} \mathbb{P}\left(\left|\frac{B_{i,1}}{\log(n)} - 1\right| > t\right) < \infty,$$

and from this, we conlude with the Borel-Cantelli lemma that $\frac{1}{\log(n)} B_{i,1} \to 1$ almost surely as $n \to \infty$.

Proof of (35): convergence of $\frac{1}{\log(n)} B_{i,2}$. We start with an integral-series comparison with the function $x \to 1/(x - X_{(i)})$ to bound $B_{i,2}$ as

$$0 \leq B_{i,2} \leq \log\left(X_{(n)} - X_{(i)}\right) - \log\left(X_{(i+r)} - X_{(i)}\right). \tag{55}$$

The first term in (55) is bounded as $\log\left(X_{(n)} - X_{(i)}\right) \leq \log\left(X_{(n)} - X_{(1)}\right)$. Let $\mu = \mathbb{E}[X]$. For any $t > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{\log(n)} \log\left(X_{(n)} - X_{(1)}\right) > t\right) &= \mathbb{P}\left(X_{(n)} - X_{(1)} > n^t\right) \\
&\leq \mathbb{P}\left(|X_{(n)} - \mu| > \frac{n^t}{2}\right) \\
&\quad + \mathbb{P}\left(|X_{(1)} - \mu| > \frac{n^t}{2}\right),
\end{aligned}$$

and

$$\mathbb{P}\left(|X_{(n)} - \mu| > \frac{n^t}{2}\right) \leq n\mathbb{P}\left(X > \mu + \frac{n^t}{2}\right) + \mathbb{P}\left(X < \mu - \frac{n^t}{2}\right)^n,$$

as well as

$$\mathbb{P}\left(|X_{(1)} - \mu| > \frac{n^t}{2}\right) \leq n\mathbb{P}\left(X < \mu - \frac{n^t}{2}\right) + \mathbb{P}\left(X > \mu + \frac{n^t}{2}\right)^n,$$

where we used the union bound twice. Since X is sub-exponential with parameters (σ, b) by assumption, it satisfies the concentration bounds stated in (20) and we deduce from the above that, for any $t > 0$ and for all n large enough,

$$\mathbb{P}\left(\frac{\log\left(X_{(n)} - X_{(1)}\right)}{\log(n)} > t\right) \leq 2n \exp\left(-\frac{n^t}{4b}\right) + 2\exp\left(-\frac{n^{1+t}}{4b}\right),$$

and since for any $t > 0$ the sum over $n \geq 1$ of the term in the right-hand side of the equation above is finite, we conclude by the Borel-Cantelli lemma that

$$\limsup_n \frac{1}{\log(n)} \log\left(X_{(n)} - X_{(1)}\right) \leq 0 \quad \text{almost surely.} \tag{56}$$

Now we bound the second term in (55). For any $t > 0$, we have

$$\mathbb{P}\left(\frac{1}{\log(n)} \log\left(\frac{1}{X_{(i+r)} - X_{(i)}}\right) > t\right) = \mathbb{P}\left(X_{(i+r)} - X_{(i)} < \frac{1}{n^t}\right). \tag{57}$$

On the event $\mathcal{E}_{n,1}$ we may expand G at $X_{(i)}$, so that there exists some $[p-\delta, p+\delta]$-valued random variable ξ such that, on the event $\mathcal{E}_{n,1}$, we have $X_{(i+r)} - X_{(i)} = G'(\xi)\left(U_{(i+r)} - U_{(i)}\right) \geq \frac{1}{\kappa_2}\left(U_{(i+r)} - U_{(i)}\right)$. Hence, for any $t > 0$,

$$\mathbb{P}\left(X_{(i+r)} - X_{(i)} < \frac{1}{n^t}\right) \leq \mathbb{P}\left(U_{(i+r)} - U_{(i)} < \frac{\kappa_2}{n^t}\right) + \mathbb{P}\left(\mathcal{E}_{n,1}^c\right). \tag{58}$$

We have $U_{(i+r)} - U_{(i)} = \mathbb{F}_n^{-1}\left(\frac{i+r}{n}\right) - \mathbb{F}_n^{-1}\left(\frac{i}{n}\right) \geq \frac{r}{n} - 2\|\mathbb{F}_n^{-1} - I\|_\infty$, as well as $\frac{1/n^t}{r/n} \to 0$ as $n \to \infty$, which we use to bound the first term in (58) as

$$\mathbb{P}\left(U_{(i+r)} - U_{(i)} < \frac{\kappa_2}{n^t}\right) \leq \mathbb{P}\left(\|\mathbb{F}_n^{-1} - I\| > \frac{r}{4n}\right),$$

for all n large enough that $\frac{\kappa_2}{n^t} \leq \frac{r}{2n}$. By Proposition 15,

$$\mathbb{P}\left(\|\mathbb{F}_n^{-1} - I\| > \frac{r}{4n}\right) \leq 2\exp\left(-\frac{r^2}{8n}\right),$$

which implies that $\sum_{n\geq 1}\mathbb{P}\left(U_{(i+r)} - U_{(i)} < \frac{\kappa_2}{n^t}\right) < \infty$ for any $t > 0$ since $\frac{r^2}{8n} > 2\log(n)$ for all n large enough. We have shown in (53) that $\sum_{n\geq 1}\mathbb{P}\left(\mathcal{E}_{n,1}^c\right) < \infty$. Consequently, with (58), we obtain that for any $t > 0$,

$$\sum_{n\geq 1}\mathbb{P}\left(X_{(i+r)} - X_{(i)} < \frac{1}{n^t}\right) < \infty,$$

and with the Borel-Cantelli lemma, this yields

$$\limsup_n \frac{1}{\log(n)}\log\left(\frac{1}{X_{(i+r)} - X_{(i)}}\right) \leq 0 \quad \text{almost surely.} \tag{59}$$

Finally, combining (56) and (59) with the bound (55), we conclude that $\frac{B_{i,2}}{\log(n)} \to 0$ almost surely as $n \to \infty$.

8.2.2 Proof of Theorem 7, statements *(ii)* and *(iii)*

We start by proving the convergence when the support admits a left endpoint (statement *(ii)*). By continuity of f over $[a, a+\epsilon)$ and the assumption that $f(a) > 0$, there exists a positive real number $\delta > 0$ such that $f(G(y))$ is bounded away from 0 over $[0, \delta]$ and we let

$$\begin{aligned}
\kappa_1 &= \inf\{f(G(y)) \,:\, 0 \leq y \leq \delta\},\\
\kappa_2 &= \sup\{f(G(y)) \,:\, 0 \leq y \leq \delta\},\\
\kappa_3 &= \sup\{|f'(G(y))| \,:\, 0 \leq y \leq \delta\}.
\end{aligned} \tag{60}$$

By (19), we have

$$\bar{H}_1 = \sum_{j=2}^{n} \frac{X_{(j)} - X_{(j-1)}}{X_{(j)} - X_{(1)}},$$

so that $\bar{H}_1 = B_1$, where B_1 is the right series B_i defined in (32) taken with $i = 1$ (the expression given there for B_i with $2 \leq i \leq n-1$ is valid for $i = 1$). The arguments used in proving the convergence of B_i apply here, with the use of the constants κ_1, κ_2 and κ_3 in (60) to deduce that $\frac{1}{\log n}B_1 \to 1$ almost surely as $n \to \infty$. In the same way, we may express $\bar{H}_n$ as $\bar{H}_n = A_1$, where A_1 is the left series defined in (32) taken with $i = 1$ to conclude that $\frac{1}{\log n}\bar{H}_n \to 1$ almost surely as $n \to \infty$, thereby proving that statement *(iii)* holds.

8.2.3 Proof of Theorem 8

We start by defining uniform versions of the constants introduced in (36). Since f is bounded away from 0 on $[x_{p_1}, x_{p_2}]$ and continuous on a neighborhood of this interval, there exists a positive real number $\delta > 0$ such that $y \mapsto f(G(y))$ is bounded away from 0 and continuous over $[p_1 - \delta, p_2 + \delta]$, and we define the following constants:

$$\begin{aligned}
\kappa_1 &= \inf\{f(G(y)) \, : \, p_1 - \delta \leq y \leq p_2 + \delta\}, \\
\kappa_2 &= \sup\{f(G(y)) \, : \, p_1 - \delta \leq y \leq p_2 + \delta\}, \\
\kappa_3 &= \sup\{|f'(G(y))| \, : \, p_1 - \delta \leq y \leq p_2 + \delta\},
\end{aligned}$$

and we let $\eta = \delta \wedge \frac{\kappa_1^3}{2\kappa_2(\kappa_3 \vee 1)}$. Let $\mathcal{I}_n = \{i \, : \, \lfloor p_1 n \rfloor \leq i \leq \lfloor p_2 n \rfloor\}$. As in (32), $\bar{H}_i$ decomposes into $\bar{H}_i = A_i + B_i$ for any $i \in \mathcal{I}_n$, and we only prove that

$$\max\left\{\left|\frac{B_i}{\log(n)} - 1\right| \, : \, i \in \mathcal{I}_n\right\} \to 0 \qquad \text{almost surely as } n \to \infty, \tag{61}$$

given that a similar convergence result may be established for A_i by using the same arguments, as in the proof of Theorem 7. Let $r := r_n = \lfloor n/\log(n) \rfloor$. For each $i \in \mathcal{I}_n$, we decompose B_i into $B_i = B_{i,1} + B_{i,2}$, where $B_{i,1} = \sum_{j=i+1}^{i+r} \frac{X_{(j)} - X_{(j-1)}}{X_{(j)} - X_{(i)}}$ and where $B_{i,2} = \sum_{j=i+r+1}^{n} \frac{X_{(j)} - X_{(j-1)}}{X_{(j)} - X_{(i)}}$, and we prove (61) by showing that

$$\frac{1}{\log(n)} \left|\max_{i \in \mathcal{I}_n} B_{i,1} - 1\right| \to 0 \quad \text{almost surely as } n \to \infty, \tag{62}$$

and that

$$\frac{1}{\log(n)} \max_{i \in \mathcal{I}_n} B_{i,2} \to 0 \quad \text{almost surely as } n \to \infty. \tag{63}$$

Letting $U_i = F(X_i)$, for $1 \leq i \leq n$, we define the following event:

$$\mathcal{E}(n) = \left[U_{(\lfloor p_1 n \rfloor)} \geq p_1 - \delta\right] \cap \left[U_{(\lfloor p_2 n \rfloor)} \leq p_2 + \delta\right] \cap \left(\cap_{i \in \mathcal{I}_n} \left[U_{(i+r)} - U_{(i)} \leq \eta\right]\right),$$

and we prove first that

$$\sum_{n \geq 1} \mathbb{P}\left(\mathcal{E}_n^c\right) < \infty. \tag{64}$$

The event $\left[U_{(\lfloor p_1 n \rfloor)} \geq p_1 - \delta\right]$ contains the event $\left[\mathbb{F}_n^{-1}\left(p_1 - \frac{1}{n}\right) > p_1 - \delta\right]$, which in turn is implied by the event $\left[\|\mathbb{F}_n^{-1} - I\|_\infty \leq \delta - \frac{1}{n}\right]$. Likewise, the event $\left[U_{(\lfloor p_2 n \rfloor)} \leq p_2 + \delta\right]$ contains the event $\left[\mathbb{F}_n^{-1}(p_2) \leq p_2 + \delta\right]$, which is implied by the event $\left[\|\mathbb{F}_n^{-1} - I\|_\infty \leq \delta\right]$. Next, for each $i \in \mathcal{I}_n$, we have $\left[U_{(i+r)} - U_{(i)} \leq \eta\right] = \left[\mathbb{F}_n^{-1}\left(\frac{i+r}{n}\right) - \mathbb{F}_n^{-1}\left(\frac{i}{n}\right) \leq \eta\right]$ which contains the event $\left[\|\mathbb{F}_n^{-1} - I\|_\infty \leq \frac{\eta - r/n}{2}\right]$. Consequently, for all n large enough that $\frac{1}{n} \leq \frac{\delta}{2}$ and $\frac{r}{n} \leq \frac{\eta}{2}$,

$$\mathbb{P}\left(\mathcal{E}(n)\right) \geq \mathbb{P}\left(\|\mathbb{F}_n^{-1} - I\|_\infty \leq \frac{\delta}{2} \wedge \frac{\eta}{4}\right), \tag{65}$$

and using Proposition 15 this yields (64).

Reproducing the steps used in proving statement (i) of Theorem 7, we obtain that (45) holds uniformly over $i \in \mathcal{I}_n$, meaning that there exists a constant $\tilde{c}_1 > 0$ depending only on f, on δ and on η such that, on the event $\mathcal{E}(n)$,

$$\max_{i \in \mathcal{I}_n} \left|B_{i,1} - \sum_{j=i+1}^{i+r} \frac{U_{(j)} - U_{(j-1)}}{U_{(j)} - U_{(i)}}\right| \leq \tilde{c}_1,$$

which implies that

$$\left|\max_{i\in\mathcal{I}_n} B_{i,1} - \max_{i\in\mathcal{I}_n}\sum_{j=i+1}^{i+r}\frac{U_{(j)}-U_{(j-1)}}{U_{(j)}-U_{(i)}}\right| \leq \tilde{c}_1.$$

Hence for any $t>0$, and for all n large enough that $\frac{\tilde{c}_1}{\log(n)} < t/2$ and that $\left|1-\frac{\log(n)}{\log(r)}\right| < t/4$,

$$\mathbb{P}\left(\max_{i\in\mathcal{I}_n}\left|\frac{1}{\log(n)}B_{i,1}-1\right|>t\right) \leq \mathbb{P}\left(\max_{i\in\mathcal{I}_n}\left|\frac{1}{\log(r)}\sum_{j=i+1}^{i+r}\frac{U_{(j)}-U_{(j-1)}}{U_{(j)}-U_{(i)}}-1\right|>\frac{t}{4}\right) + \mathbb{P}(\mathcal{E}_n^c). \tag{66}$$

Using the representation of the uniform order statistics in terms of exponential variables given in (49), it follows that jointly,

$$\begin{aligned}&\left\{\sum_{j=i+1}^{i+r}\frac{U_{(j)}-U_{(j-1)}}{U_{(j)}-U_{(i)}} \,:\, i\in\mathcal{I}_n\right\}\\ &\quad\overset{\mathcal{L}}{=}\left\{\sum_{j=i+1}^{i+r}\frac{\nu_j}{\nu_{i+1}+\cdots+\nu_j} \,:\, i\in\mathcal{I}_n\right\}\\ &\quad\overset{\mathcal{L}}{=}\left\{\sum_{j=i+1}^{i+r}\frac{\nu_j}{\nu_{i+1}+\cdots+\nu_j} \,:\, 0\leq i\leq \lfloor(p_2-p_1)n\rfloor\right\}.\end{aligned}$$

Applying Lemma 14 together with the union bound, we obtain the bound

$$\begin{aligned}&\mathbb{P}\left(\max_{i\in\mathcal{I}_n}\left|\frac{1}{\log(r)}\sum_{j=i+1}^{i+r}\frac{U_{(j)}-U_{(j-1)}}{U_{(j)}-U_{(i)}}-1\right|>\frac{t}{4}\right)\\ &\quad\leq 2n\exp\left(-\frac{t\log(r)}{16}\log\left(1+\frac{3t\log(r)}{4\pi^2}\right)\right),\end{aligned}$$

which holds for any $t>0$ and for all n large enough, and for any $t>0$, this bound is in turn bounded by $\frac{2}{n^2}$ for all n large enough that $\log(r)>\log(n)/2$ and that $\frac{t}{32}\log\left(1+\frac{3t\log(n)}{8\pi^2}\right)-1>2$. Using this in (66) together with (64), and applying the Borel-Cantelli lemma, we deduce that $\max_{i\in\mathcal{I}_n}\left|\frac{1}{\log(n)}B_{i,1}-1\right|\to 0$ almost surely as $n\to\infty$ which proves (62).

To prove (63), we start with the bound

$$0\leq\max_{i\in\mathcal{I}_n}B_{i,2}\leq\log\left(X_{(n)}-X_{(1)}\right)+\log\left(\max_{i\in\mathcal{I}_n}\frac{1}{X_{(i+r)}-X_{(i)}}\right), \tag{67}$$

which is a uniform version of (55) over $\mathcal{I}_n$, and where we used the monotony of the logarithm function. We have proved in (56) that $\frac{1}{\log(n)}\log\left(X_{(n)}-X_{(1)}\right)\to 0$ almost surely as $n\to\infty$ so we only need to prove that the limit superior of the last term in (67) is bounded by 0 with probability one. Proceeding as in the proof of Theorem 7, we obtain that, for any $t>0$ and for all n large enough,

$$\mathbb{P}\left(\frac{1}{\log(n)}\log\left(\max_{i\in\mathcal{I}_n}\frac{1}{X_{(i+r)}-X_{(i)}}\right)>t\right)\leq\mathbb{P}\left(\min_{i\in\mathcal{I}_n}\left(U_{(i+r)}-U_{(i)}\right)<\frac{\kappa_2}{n^t}\right) + \mathbb{P}\left(\mathcal{E}_n^c\right). \tag{68}$$

We have $U_{(i+r)}-U_{(i)}\geq\frac{r}{n}-2\|\mathbb{F}_n^{-1}-I\|$ for all $i\in\mathcal{I}_n$, so that, for any $t>0$, and for all n large enough that $\frac{\kappa_2}{n^t}\leq\frac{r}{2n}$,

$$\mathbb{P}\left(\min_{i\in\mathcal{I}_n}\left(U_{(i+r)}-U_{(i)}\right)<\frac{\kappa_2}{n^t}\right)\leq n\mathbb{P}\left(\|\mathbb{F}_n^{-1}-I\|>\frac{r}{8n}\right)\leq 2n\exp\left(-\frac{r^2}{4n}\right),$$

where we used the union bound and then Proposition 15 in the last inequality. Therefore, $\sum_{n\geq 1}\mathbb{P}\left(\min_{i\in\mathcal{J}_n}\left(U_{(i+r)}-U_{(i)}\right)<\frac{\kappa_2}{n^t}\right)<\infty$ for any $t>0$ and with (64), (68) and the Borel-Cantelli lemma, we obtain that

$$\limsup_n \frac{1}{\log(n)}\log\left(\max_{i\in\mathcal{J}_n}\frac{1}{X_{(i+r)}-X_{(i)}}\right)\leq 0 \quad \text{almost surely.}$$

Then using the bounds in (67), we conlude that $\frac{1}{\log(n)}\max_{i\in\mathcal{J}_n}B_{i,2}\to 0$ almost surely as $n\to\infty$, which proves (63).

8.3 Asymptotics in a fixed design

In this section we prove Theorem 10 and Theorem 11. We recall that $\mathcal{H}_\ell$ denotes the ℓ^{th} harmonic number defined by $\mathcal{H}_\ell=\sum_{k=1}^{\ell}\frac{1}{k}$.

8.3.1 Proof of Theorem 10

We first prove the pointwise statement. When $d=1$, applying Theorem 5 leads to

$$\begin{cases}\mathbb{E}[h_{\mathcal{T}}(x_1)]=\mathbb{E}[h_{\mathcal{T}}(x_n)]=\mathcal{H}_{n-1}, & \text{and}\\ \mathbb{E}[h_{\mathcal{T}}(x_i)]=\mathcal{H}_{i-1}+\mathcal{H}_{n-i}, & \text{for } 2\leq i\leq n-1,\end{cases} \tag{69}$$

where $x_i=(i-1)/(n-1)$ for $i\in\{1,\dots,n\}$. By Proposition 6, we have $\mathbb{E}[h_{\mathcal{T}}(0)]=\mathbb{E}[h_{\mathcal{T}}(x_1)]$ and $\mathbb{E}[h_{\mathcal{T}}(1)]=\mathbb{E}[h_{\mathcal{T}}(x_n)]$, and for any $0<x<1$, for all n large enough, $\mathbb{E}[h_{\mathcal{T}}(x)]$ is a convex combination of $\mathbb{E}[h_{\mathcal{T}}(x_{i(x)})]$ and $\mathbb{E}[h_{\mathcal{T}}(x_{i(x)+1})]$, with $i(x)=1+\lfloor(n-1)x\rfloor$. The result then follows using the inequalities $\log(\ell+1)\leq\mathcal{H}_\ell\leq 1+\log(\ell)$, for any $\ell\geq 1$.

We now assume that $d\geq 2$. Let $\mathbf{i}=(i_1,\dots,i_d)\in\{1,\dots,n\}^d$, and let $x_{\mathbf{i}}=(x_{\mathbf{i}}^{(1)},\dots,x_{\mathbf{i}}^{(d)})$, where $x_{\mathbf{i}}^{(j)}=i_j/(n-1)$, for any $j\in\{1,\dots,d\}$. For $\ell\in\{1,\dots,d\}$, we denote by $P^{(\ell)}:\mathbb{R}^d\to\mathbb{R}$ the projection operator acting as $P^{(\ell)}(x)=x^{(\ell)}$.

We recall first that during the growth of $(\mathcal{T},\pi_{\mathcal{T}})$ with Algorithm 1, if j is selected as the split component to partition $\pi_{\mathcal{T}}(v)$ at a node v, then the split value τ is drawn uniformly between the minimal and maximal value of $P^{(j)}\left(\pi_{\mathcal{T}}(v)\cap\mathcal{D}_n\right)$. In this case for each child $v\eta$ of v, with $\eta\in\{0,1\}$, we have $P^{(\ell)}\left(\pi_{\mathcal{T}}(v\eta)\cap\mathcal{D}_n\right)=P^{(\ell)}\left(\pi_{\mathcal{T}}(v)\cap\mathcal{D}_n\right)$ for any $\ell\neq j$, due to the fact that the points in $\mathcal{D}_n$ are arranged as a regular grid, thus leaving unchanged the support of the distribution of a subsequent split value along a component different from j, as well as the number of distinct points in each set $P^{(\ell)}\left(\pi_{\mathcal{T}}(v\eta)\cap\mathcal{D}_n\right)$ for $\ell\neq j$. It follows from this that $h_{\mathcal{T}}(x_{\mathbf{i}})$ is distributed according to $h_{\mathcal{T}_1}(x_{\mathbf{i}}^{(1)})+\cdots+h_{\mathcal{T}_d}(x_{\mathbf{i}}^{(d)})$, where $\mathcal{T}_1,\dots,\mathcal{T}_d$ denote d independent univariate random isolation trees defined using Algorithm 1 using the set $\mathcal{D}_n^{(1)}:=\{(i-1)/(n-1):1\leq i\leq n\}$.

Let $\mathcal{J}_0=\left\{1\leq j\leq d:i_j\in\{2,\dots,n-1\}\right\}$ and let $\mathcal{J}_1=\{1,\dots,d\}\setminus\mathcal{J}_0$. Note that $x_{\mathbf{i}}$ is an interior point of $[0,1]^d$ if all the components of $\mathbf{i}$ are in $\mathcal{J}_0$ and a boundary point otherwise. Using (69), we deduce that

$$\mathbb{E}[h_{\mathcal{T}}(x_{\mathbf{i}})]=\sum_{j\in\mathcal{J}_0}\left(\mathcal{H}_{i_j-1}+\mathcal{H}_{n-i_j}\right)+(\#\mathcal{J}_1)\,\mathcal{H}_{n-1}, \tag{70}$$

where we use the convention that a sum over an empty set is equal to 0.

Let x be an interior point of $[0,1]^d$. For any $j\in\{1,\dots,d\}$, let $i_j(x)=1+\lfloor(n-1)x^{(j)}\rfloor$, and let $\mathbf{i}(x)=(i_1(x),\dots,i_d(x))$. Then x belongs to the cube with vertices $\{x_{\mathbf{i}}:\mathbf{i}\in\mathbf{I}(x)\}$ where $\mathbf{I}(x):=\left\{\mathbf{i}(x)+\delta:\delta\in\{0,1\}^d\right\}$. By Proposition 6, the value of $\mathbb{E}[h_{\mathcal{T}}(x)]$ is a convex combination (with coefficients depending on x) of $\{\mathbb{E}[h_{\mathcal{T}}(x_{\mathbf{i}})]:\mathbf{i}\in\mathbf{I}(x)\}$. Since x is an interior point, $\mathcal{J}_0=\{1,\dots,d\}$ and $\mathcal{J}_1=\emptyset$ for each n so that $i_j(x)/n\to x$ as $n\to\infty$ for each $j\in\{1,\dots,d\}$.

Using this, together with the fact that $\mathcal{H}_\ell/\log(\ell) \to 1$ as $\ell \to \infty$ yields $\mathbb{E}[h_{\mathcal{T}}(x_{\mathbf{i}})]/\log(n) \to 2d$ as $n \to \infty$. This proves the result when $k = d$ since $\mathcal{F}_d$ contains only $[0,1]^d$.

Suppose now that $x \in \mathring{F}_k$ for some face $F_k \in \mathcal{F}$, with $0 \le k < d$. We argue as above inside the face F_k. Let $\mathbf{i}(x) = (i_1(x), \dots, i_d(x))$ where $i_j(x) = 1 + \lfloor (n-1)x^{(j)} \rfloor$ if $j \in \mathcal{J}_0$ and where $i_j(x) = i_j$ when $j \in \mathcal{J}_1$. Then x belongs to the k-dimensional cube with vertices $\{x_{\mathbf{i}} : \mathbf{i} \in \mathbf{I}(x)\}$ where $\mathbf{I}(x) = \{\mathbf{i}(x)+\delta : \delta \in \{0,1\}\}$. Notice that this cube is included in the interior of F_k, which is also a k-dimensional cube. Since $\#\mathcal{J}_0 = k$ and $\#\mathcal{J}_1 = d-k$, using Proposition 6 together with the relation $\mathcal{H}_\ell/\log(\ell) \to 1$ as $\ell \to \infty$, it follows that $\mathbb{E}[h_{\mathcal{T}}(x_{\mathbf{i}})]/\log(n) \to 2k + (d-k) = d + k$ as $n \to \infty$, which yields the desired result.

To prove that the convergence is uniform over any closed subset contained in the interior of $[0,1]^d$, it suffices to show that this holds in dimension $d = 1$ and to argue as above. Let $0 < x_1 < x_2 < 1$ and let $i_1 = 1 + \lfloor (n-1)x_1 \rfloor$ and $i_2 = 2 + \lfloor (n-1)x_2 \rfloor$. By (70), we have $\mathbb{E}[h_{\mathcal{T}}(x_i)] = \mathcal{H}_{i-1} + \mathcal{H}_{n-i}$ for all $i_1 \le i \le i_2$. Using the inequalities $\log(\ell+1) \le \mathcal{H}_\ell \le 1 + \log(\ell)$ for any $\ell \ge 1$ yields

$$\sup_{i_1 \le i \le i_2} \left| \frac{1}{\log(n)} \left(\mathcal{H}_{i-1} + \mathcal{H}_{n-i} \right) - 2 \right| \to 0, \quad \text{as } n \to \infty. \tag{71}$$

Using this together with Proposition 6 we conclude that

$$\sup_{x_1 \le x \le x_2} \left| \frac{1}{\log(n)} \mathbb{E}[h_{\mathcal{T}}(x)] - 2 \right| \to 0 \quad \text{as } n \to \infty.$$

8.4 Proof of Theorem 11

For $k \in \{1, \dots, K\}$, let $N_k = n_1 + \cdots + n_k$ and let $N_0 = 0$. Using Theorem 5, we first prove that for each $k \in \{1, \dots, K\}$ we have

$$\mathbb{E}[h_{\mathcal{T}}(x_i)] = \begin{cases} \mathcal{H}_{n_k - 1} + R_i & \text{if } i = N_{k-1} + 1, \\ \mathcal{H}_{i - N_{k-1} - 1} + \mathcal{H}_{N_k - i} + R_i & \text{if } N_{k-1} + 2 \le i \le N_k - 1, \\ \mathcal{H}_{n_k - 1} + R_i & \text{if } i = N_k, \end{cases} \tag{72}$$

where the R_i's are remainder terms satisfying

$$\sup_{n \ge 1} \sup_{1 \le i \le n} |R_i| \le C_R, \tag{73}$$

for a constant $C_R > 0$ depending only on the interval lengths $L_1, \dots, L_K$ and the gaps $\delta_1, \dots, \delta_{K-1}$. Notice that the expressions involving harmonic numbers in (72) corresponds those obtained in (69) for the case of n_k equally spaced points. The remainder terms account for the fact that the summation range in the series in Theorem 5 extends over all the intervals. By Proposition 6, the value of $\mathbb{E}[h_{\mathcal{T}}(x)]$ at any $x \in \mathbb{R}$ is obtained by linear interpolation of $\{(x_i, h_{\mathcal{T}}(x_i)) : 1 \le i \le n\}$. Therefore if (72) and (73) hold, then using Proposition 6 together with the fact that $\log(n_k)/\log(n) \to 1$ since $n_k/n \to \alpha_k$ as $n \to \infty$ for each $k \in \{1, \dots, K\}$ and the fact that $\mathcal{H}_\ell/\log(\ell) \to 1$ as $\ell \to \infty$ leads to the desired result. The conlusion that the convergence is uniform then follows by the same arguments as those used in proving (71).

There remains to prove (72) and (73). We do so for the first interval only, the reasoning being identical for the other intervals. Recall that the configuration is scaled to extend over $[0,1]$, so that $x_1 = 0$ and $x_n = 1$. For the first boundary point x_1 of $\mathcal{I}_1$, by Theorem 5, we have

$$\mathbb{E}[h_{\mathcal{T}}(x_1)] = \sum_{j=1}^{n-1} \frac{x_{j+1} - x_j}{x_{j+1} - x_1},$$

which decomposes into

$$\mathbb{E}[h_{\mathcal{T}}(x_1)] = \sum_{j=1}^{n_1-1} \frac{x_{j+1}-x_j}{x_{j+1}-x_1} + \sum_{j=n_1}^{n-1} \frac{x_{j+1}-x_j}{x_{j+1}-x_1}. \tag{74}$$

The points in $\mathcal{I}_1$ being equally spaced, the first sum in the right hand side of (74) is equal to $\sum_{j=1}^{n_1-1} \frac{1}{j} = \mathcal{H}_{n_1-1}$, and by an integral-series comparison, the second sum in (74) is bounded from above by $\log(x_n) - \log(x_{n_1}) = \log(1/L_1)$. Therefore $\mathbb{E}[h_{\mathcal{T}}(x_1)] = \mathcal{H}_{n_1-1} + R_1$ where R_1 satisfies $\sup_{n\geq 1} |R_1| \leq \log(1/L_1)$.

For $2 \leq i \leq n_1 - 1$, by Theorem 5, we have

$$\mathbb{E}[h_{\mathcal{T}}(x_i)] = \sum_{j=1}^{i-1} \frac{x_{j+1}-x_j}{x_i - x_j} + \sum_{j=i+1}^{n} \frac{x_j - x_{j-1}}{x_j - x_i}. \tag{75}$$

The first sum in the right hand side of (75) is equal to $\mathcal{H}_{i-1}$, and the second sum decomposes into

$$\sum_{j=i+1}^{n} \frac{x_j - x_{j-1}}{x_j - x_i} = \sum_{j=i+1}^{n_1} \frac{x_j - x_{j-1}}{x_j - x_i} + \frac{x_{n_1+1} - x_{n_1}}{x_{n_1+1} - x_i} + \sum_{j=n_1+2}^{n} \frac{x_j - x_{j-1}}{x_j - x_i}. \tag{76}$$

The first term in the right hand side of (76) is equal to $\mathcal{H}_{n_1-i}$, while the second term is smaller than 1, and by an integral-series comparison, the last sum is bounded from above by $\log(x_n - x_i) - \log(x_{n_1+1} - x_i) \leq \log(1/\delta_1)$. This yields $\mathbb{E}[h_{\mathcal{T}}(x_i)] = \mathcal{H}_i + \mathcal{H}_{n_1-i} + R_i$ where R_i satisfies $\sup_{n\geq 1} \sup_{2\leq i\leq n_1-1} |R_i| \leq 1 + \log(1/\delta_1)$.

At last, for $i = n_1$, by Theorem 5, we have

$$\mathbb{E}[h_{\mathcal{T}}(x_{n_1})] = \sum_{j=1}^{n_1-1} \frac{x_{j+1}-x_j}{x_{n_1} - x_j} + \sum_{j=n_1+1}^{n} \frac{x_j - x_{j-1}}{x_j - x_{n_1}} = \mathcal{H}_{n_1-1} + \sum_{j=n_1+1}^{n} \frac{x_j - x_{j-1}}{x_j - x_{n_1}},$$

and the last sum is bounded by $1 + \log(x_n - x_{n_1+1}) - \log(x_{n_1+1} - x_{n_1}) \leq 1 + \log(1/\delta_1)$, which yields

$$\mathbb{E}[h_{\mathcal{T}}(x_{n_1})] = \mathcal{H}_{n_1-1} + R_{n_1},$$

where $\sup_{n\geq 1} |R_{n_1}| \leq 1 + \log(1/\delta_1)$. Hence we have shown that $\mathbb{E}[h_{\mathcal{T}}(x_i)] = \mathcal{H}_{i-1} + \mathcal{H}_{n_1-1} + R_i$ for any $1 \leq i \leq n_1$ where $\{R_i : 1 \leq i \leq n_1\}$ satisfy $\sup_{n\geq 1} \sup_{1\leq i\leq n_1} |R_i| \leq \log(1/L_1) \vee (1 + \log(1/\delta_1))$. Reasoning along the same lines leads to similar bounds for points in the other intervals and this proves (72) and (73).

A Auxiliary results

In this appendix we collect auxiliary results. This includes three technical Lemmas (Section A.1), a concentration inequality that relate to the uniform empirical process (Section A.2), and proofs for the bounds on the average heights for the configurations of points studied in Section 5 (Section A.3).

A.1 Technical Lemmas

Lemma 12. *Let F_0 be an absolutely continuous cumulative distribution function over $\mathbb{R}$. Given n strictly positive real numbers $w_1, \dots, w_n$, let $Z_1, \dots, Z_n$ be n independent random variables where Z_i has distribution $F_0^{w_i}$ for any $1 \leq i \leq n$. Let $i \in \{1, \dots, n\}$ and let $\mathcal{K}$ be a subset of $\{1, \dots, n\}$ containing i. Then*

$$\mathbb{P}\left(Z_i \geq \max_{k\in\mathcal{K}} Z_k\right) = \frac{w_i}{\sum_{k\in\mathcal{K}} w_k}.$$

Proof. We have

$$\begin{aligned}\mathbb{P}\left(Z_i \geq \max_{k\in\mathcal{K}} Z_k\right) &= \mathbb{E}\left[\prod_{k\in\mathcal{K}\setminus\{i\}} \mathbb{P}\left(Z_k \leq Z_i | Z_i\right)\right] \\ &= \mathbb{E}\left[F_0(Z_i)^{\sum_{k\in\mathcal{K}\setminus\{i\}} w_k}\right] \\ &= \int_{\mathbb{R}} w_i F_0'(z) F_0(z)^{\sum_{k\in\mathcal{K}} w_k - 1} dz \\ &= \frac{w_i}{\sum_{k\in\mathcal{K}} w_k}\end{aligned}$$

□

The following Lemma states an independence property for the records associated with the random variables $Z_1, \ldots, Z_n$ of Lemma 12. This property is well known when the sequence of random variables is IID.

Lemma 13. *In the setting of Lemma 12, let $B_i = \left[Z_i \geq \max_{1\leq\ell\leq i} Z_\ell\right]$, for $i = 1, \ldots, n$. Then the events $B_1, \ldots, B_n$ are independent.*

Proof. In this proof, we use the convention that a sum over an empty set of indexes is equal to 0, and that a product over an empty set of indexes is equal to 1.

Let $2 \leq k \leq n$ be an integer and let $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ be k integers. Let $\mathcal{I}_\ell = \{1, \ldots, i_\ell\}$ for $1 \leq \ell \leq k$. Let $\mathcal{J}_\ell = \mathcal{I}_\ell \setminus \mathcal{I}_{\ell-1}$ for $2 \leq \ell \leq k$ and set $\mathcal{J}_1 = \mathcal{I}_1$. We have

$$\mathbb{P}\left(B_{i_1} \cap \cdots \cap B_{i_k}\right) = \mathbb{P}\left(\left[Z_{i_1} < Z_{i_2} < \cdots < Z_{i_k}\right] \cap A_1 \cap \cdots \cap A_k\right),$$

where $A_\ell = \cap_{i\in\mathcal{J}_\ell} \left[Z_i \leq Z_{i_\ell}\right]$, for $1 \leq \ell \leq k$. We have

$$\begin{aligned}&\mathbb{P}\left(\left[Z_{i_1} < Z_{i_2} < \cdots < Z_{i_k}\right] \cap A_1 \cap \cdots \cap A_k\right) \\ &\quad = \mathbb{E}\left[\mathbf{1}\left\{Z_{i_1} < Z_{i_2} < \cdots < Z_{i_k}\right\} \mathbb{P}\left(A_1 \cap \cdots \cap A_k | Z_{i_1}, \ldots, Z_{i_k}\right)\right] \\ &\quad = \mathbb{E}\left[\mathbf{1}\left\{Z_{i_1} < Z_{i_2} < \cdots < Z_{i_k}\right\} \prod_{\ell=1}^{k} F_0(Z_\ell)^{\sum_{j\in\mathcal{J}_\ell\setminus\{i_\ell\}} w_j}\right]\end{aligned}$$

By conditioning on $Z_{i_2}, \ldots, Z_{i_k}$, we obtain that

$$\begin{aligned}&\mathbb{P}\left(\left[Z_{i_1} < Z_{i_2} < \cdots < Z_{i_k}\right] \cap A_1 \cap \cdots \cap A_k\right) \\ &\quad = \mathbb{E}\Big[\mathbf{1}\{Z_{i_2} < \cdots < Z_{i_k}\} \prod_{\ell=2}^{k} F_0(Z_\ell)^{\sum_{j\in\mathcal{J}_\ell\setminus\{i_l\}} w_j} \\ &\qquad \times \mathbb{E}\left[\mathbf{1}\{Z_{i_1} < Z_{i_2}\} F_0(Z_{i_1})^{\sum_{j\in\mathcal{J}_1\setminus\{i_1\}} w_j} | Z_{i_2}, \ldots, Z_{i_k}\right]\Big]. \qquad (77)\end{aligned}$$

We have

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}\{Z_{i_1} < Z_{i_2}\} F_0(Z_{i_1})^{\sum_{j\in\mathcal{J}_1\setminus\{i_1\}} w_j} | Z_{i_2}, \ldots, Z_{i_k}\right] \\ &= \int_{-\infty}^{Z_{i_2}} F_0(z)^{\sum_{j\in\mathcal{J}_1\setminus\{i_1\}} w_j} w_{i_1} F_0'(z) F_0(z)^{w_{i_1}-1} dz \\ &= \frac{w_{i_1}}{\sum_{j\in\mathcal{J}_1} w_j} F_0(Z_{i_2})^{\sum_{j=1}^{i_1} w_j}. \qquad (78)\end{aligned}$$

Reporting (78) in (77) yields

$$\begin{aligned}&\mathbb{P}\left(\left[Z_{i_1} < Z_{i_2} < \cdots < Z_{i_k}\right] \cap A_1 \cap \cdots \cap A_k\right)\\&= \mathbb{P}(B_{i_1})\mathbb{E}\left[\mathbf{1}\{Z_{i_2} < \cdots < Z_{i_k}\}F_0(Z_{i_2})^{\sum_{j=1}^{i_2-1} w_j}\prod_{\ell=3}^{k} F_0(Z_\ell)^{\sum_{j\in\mathcal{J}_\ell\setminus\{i_\ell\}} w_j}\right],\end{aligned}$$

where we used the fact that $\mathbb{P}(B_{i_1}) = \frac{w_{i_1}}{\sum_{j\in\mathcal{J}_1} w_j}$ by Lemma 12 together with the fact that $\mathcal{J}_1 = \mathcal{I}_1$. Iterating in the same way, we deduce that

$$\mathbb{P}\left(B_{i_1} \cap \cdots \cap B_{i_k}\right) = \mathbb{P}(B_{i_1}) \times \cdots \times \mathbb{P}(B_{i_k}).$$

□

The following lemma gives a concentration bound on a sum that arises in the proofs of Theorem 7 and Theorem 8 when representing the uniform order statistics in terms of exponential random variables. We recall that $\mathcal{H}_n$ denotes the n^{th} harmonic number defined by $\mathcal{H}_n = \sum_{i=1}^n \frac{1}{n}$ and that it satisfies $\frac{\mathcal{H}_n}{\log(n)} \to 1$ as $n \to \infty$.

Lemma 14. *Let $(\nu_i)_{(i\geq 1)}$ be a sequence of independent random variables and identically distributed according to an exponential distribution with mean equal to 1. Let $S_n = \nu_1 + \cdots + \nu_n$. For any $t > 0$, and for all n large enough that $|\mathcal{H}_n/\log(n) - 1| \leq t/2$,*

$$\mathbb{P}\left(\left|\frac{1}{\log(n)}\sum_{i=1}^{n}\frac{\nu_i}{S_i} - 1\right| > t\right) \leq 2\exp\left(-\frac{t\log(n)}{4}\log\left(1 + \frac{3t\log(n)}{\pi^2}\right)\right).$$

Proof. For any $i \geq 1$, let $Y_i = \frac{\nu_i}{S_i}$, and note that $Y_i = 1 - \frac{S_{i-1}}{S_i}$ for any $i \geq 2$. We first prove the somewhat surprising fact that $Y_2, \ldots, Y_n$ are independent. Since the random variables $\nu_1, \ldots, \nu_n$ are independent and identically distributed according to an exponential distribution with mean equal to 1, the distribution of the random vector $(S_1, \ldots, S_n)$ admits a probability density function $f_{S,n}$ defined by

$$f_{S,n}(s_1, \ldots, s_n) = e^{-s_n}\mathbf{1}\{(s_1, \ldots, s_n) \in \mathcal{D}_{S,n}\},$$

where $\mathcal{D}_{S,n} = \{(s_1, \ldots, s_n) \in \mathbb{R}^n : 0 \leq s_1 \leq s_2 \leq \cdots \leq s_n\}$. Consider the transformation $\Phi : \mathcal{D}_{S,n} \to \mathbb{R}_+ \times [0,1]^{n-1}$ defined by $\Phi(s_1, \ldots, s_n) = \left(s_1, 1 - \frac{s_1}{s_2}, \ldots, 1 - \frac{s_{n-1}}{s_n}\right)$. Its inverse function is defined over $\mathbb{R}_+ \times [0,1]^{n-1}$ by

$$\Phi^{-1}(y_1, \ldots, y_n) = \left(y_1, \frac{y_1}{1-y_2}, \frac{y_1}{(1-y_2)(1-y_3)}, \cdots, \frac{y_1}{(1-y_2)\ldots(1-y_n)}\right),$$

with Jacobian equal to $\frac{y_1^{n-1}}{\prod_{k=2}^n (1-y_k)^{n+2-k}}$. Then $(\nu_1, Y_2, \ldots, Y_n) = \Phi(S_1, \ldots, S_n)$ so that the random vector $(\nu_1, Y_2, \ldots, Y_n)$ admits a probability density function $f_{(\nu_1, Y_2, \ldots, Y_n)}$ given by

$$\begin{aligned}f_{(\nu_1, Y_2, \ldots, Y_n)}(u, y_2, \ldots, y_n) = &\frac{u^{n-1}}{\prod_{k=2}^n (1-y_k)^{n+2-k}}\exp\left(-\frac{u}{\prod_{k=2}^n (1-y_k)}\right)\\&\times \mathbf{1}\left\{(u, y_2, \ldots, y_n) \in \mathbb{R}_+ \times [0,1]^{n-1}\right\},\end{aligned}$$

from which we deduce the probability density function $f_{(Y_2, \ldots, Y_n)}$ of $(Y_2, \ldots, Y_n)$ which is expressed as

$$f_{(Y_2, \ldots, Y_n)}(y_2, \ldots, y_n) = (n-1)!\prod_{k=2}^{n}(1-y_k)^{k-2}\mathbf{1}\left\{(y_2, \ldots, y_n) \in [0,1]^{n-1}\right\}.$$

Therefore the variables $Y_2, \dots, Y_n$ are independent.

Also, for each $i \geq 1$, Y_i follows a Beta distribution $\mathrm{Beta}(1, i-1)$, since $Y_i = \frac{\nu_i}{S_i}$ by definition, and so $\mathbb{E}[Y_i] = \frac{1}{i}$ and $\mathrm{Var}(Y_i) = \frac{i-1}{i^2(i+1)} \leq \frac{1}{i^2}$. Hence for any $n \geq 2$, we have

$$\sum_{i=1}^{n} \mathbb{E}\left[\frac{\nu_i}{S_i}\right] = \mathcal{H}_n \quad \text{and} \quad \sigma_n^2 := \sum_{i=1}^{n} \mathrm{Var}\left(\frac{\nu_i}{S_i}\right) \leq \sum_{i=2}^{n} \frac{1}{i^2} \leq \frac{\pi^2}{6}. \tag{79}$$

For any $t > 0$ and for all n large enough that $|\mathcal{H}_n / \log(n) - 1| \leq t/2$, we have

$$\mathbb{P}\left(\left|\frac{1}{\log(n)} \sum_{i=1}^{n} \frac{\nu_i}{S_i} - 1\right| > t\right) \leq \mathbb{P}\left(\left|\sum_{i=1}^{n} \frac{\nu_i}{S_i} - \mathcal{H}_n\right| > \frac{t \log(n)}{2}\right). \tag{80}$$

Using Bennett's inequality, for any $t > 0$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \frac{\nu_i}{S_i} - \mathcal{H}_n\right| > \frac{t \log(n)}{2}\right) \leq 2 \exp\left(-\sigma_n^2 h\left(\frac{t \log(n)}{2\sigma_n^2}\right)\right),$$

where h is the function defined by $h(u) = (1+u)\log(1+u) - u$, and where we used the facts that $0 \leq \frac{\nu_i}{S_i} \leq 1$ almost surely for all $i \geq 1$, that $\nu_1/S_1 = 1$, and that the variables $\nu_2/S_2, \dots, \nu_n/S_n$ are independent. Using the inequality $h(u) \geq \frac{1}{2} u \log(1+u)$ for any $u \geq 0$ and the bound $\sigma_n^2 \leq \pi^2/6$ in (79), this leads to

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \frac{\nu_i}{S_i} - \mathcal{H}_n\right| > \frac{t \log(n)}{2}\right) \leq 2 \exp\left(-\frac{t \log(n)}{4} \log\left(1 + \frac{3t \log(n)}{\pi^2}\right)\right),$$

which, combined with (80), yields the desired result. □

A.2 The uniform empirical quantile process

The following Proposition gives an exponential inequality for the uniform empirical process and the uniform quantile process, using the DKW inequality (Dvoretzky et al., 1956) with the tight constant due to Massart (1990).

Proposition 15. *Let $U_1, \dots, U_n$ be IID random variables distributed according to a uniform distribution over $[0,1]$. Let $\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{U_i \leq t\}$ for $0 \leq t \leq 1$ and $\mathbb{F}_n^{-1}(p) = \inf\{y : \mathbb{F}_n(y) \geq p\}$, for $0 \leq p \leq 1$. Then for any $\lambda > 0$,*

$$\mathbb{P}\left(\sup_{p \in [0,1]} \sqrt{n} |\mathbb{F}_n^{-1}(p) - p| > \lambda\right) = \mathbb{P}\left(\sup_{t \in [0,1]} \sqrt{n} |\mathbb{F}_n(t) - t| > \lambda\right) \leq 2\exp(-2\lambda^2). \tag{81}$$

A.3 Average heights of the deterministic configurations

In this section we prove the inequalities on the average heights of each configuration of points studied in Section 5. For clarity, we start each paragraph by recalling the definition of each configuration of points.

Proof of (15). The n points are such that $x_1 = 0$ and $1 - \epsilon = x_2 < x_3 < \cdots < x_{n-1} < x_n = 1$. Using Theorem 5, we have

$$\mathbb{E}[h_{\mathcal{T}}(x_1)] = 1 + \sum_{j=3}^{n} \frac{x_j - x_{j-1}}{x_j - x_1} \leq 1 + \sum_{j=3}^{n} \frac{x_j - x_{j-1}}{x_2 - x_1} = 1 + \frac{x_n - x_2}{x_2 - x_1} = 1 + \frac{\epsilon}{1-\epsilon}.$$

For $2 \leq i \leq n-1$, we note that each of the two sums given in Theorem 5 for the expression of $\mathbb{E}[h_{\mathcal{T}}(x_i)]$ contains one term equal to 1, implying that $\mathbb{E}[h_{\mathcal{T}}(x_i)] \geq 2$. At last,

$$\mathbb{E}[h_{\mathcal{T}}(x_n)] = 1 + \sum_{j=1}^{n-2} \frac{x_{j+1} - x_j}{x_n - x_j} \geq 1 + \sum_{j=1}^{n-2} (x_{j+1} - x_j) \geq 2 - \epsilon.$$

Therefore $\mathbb{E}[h_{\mathcal{T}}(x_i)] \geq 2 - \epsilon$ for any $2 \leq i \leq n$.

Proof of (16). The point configuration is such that $x_1 = 0$ and that $x_2, \dots, x_n$ extend uniformly over $[1-\epsilon, 1]$, so that $x_{j+1} - x_j = \epsilon/(n-2)$ for any $2 \leq j \leq n-1$. Applying Theorem 5, we obtain that

$$\mathbb{E}[h_{\mathcal{T}}(x_1)] = 1 + \sum_{j=3}^{n} \frac{\epsilon/(n-2)}{1-\epsilon+(j-2)\epsilon/(n-2)} \leq 1 + \frac{\epsilon}{1-\epsilon},$$

and that

$$\mathbb{E}[h_{\mathcal{T}}(x_2)] = 1 + \sum_{j=3}^{n} \frac{x_j - x_{j-1}}{x_j - x_2} = 1 + \sum_{j=3}^{n} \frac{1}{j-2} = 1 + \mathcal{H}_{n-2}.$$

For $3 \leq i \leq n-1$, we have

$$\mathbb{E}[h_{\mathcal{T}}(x_i)] \geq \sum_{j=2}^{i-1} \frac{x_{j+1} - x_j}{x_i - x_j} + \sum_{j=i+1}^{n} \frac{x_j - x_{j-1}}{x_j - x_i} = \mathcal{H}_{i-2} + \mathcal{H}_{n-i}.$$

At last, we have

$$\mathbb{E}[h_{\mathcal{T}}(x_n)] = 1 - \epsilon + \sum_{j=2}^{n-1} \frac{x_{j+1} - x_j}{x_n - x_j} = 1 - \epsilon + \mathcal{H}_{n-2}.$$

Using the inequality $\mathcal{H}_n \geq \log(n+1)$, simple calculations leads to $\mathcal{H}_{i-2} + \mathcal{H}_{n-i} \geq \log((i-1)(n-i+1)) \geq \log(2n-4) \geq \log(n-1)$ for any $2 \leq i \leq n-1$. Therefore we conclude that $\mathbb{E}[h_{\mathcal{T}}(x_i)] \geq \log(n-1)$ for any $2 \leq i \leq n$.

Proof of (17). In this configuration the points are defined by the recursion $x_{j+1} = 1 - \epsilon(1 - x_j)$ with $x_1 = 0$, so that $x_j = 1 - \epsilon^j$ for any $1 \leq j \leq n$, and that $x_{j+1} - x_j = \epsilon^{j-1} - \epsilon^j$, for any $1 \leq j \leq n-1$. Applying Theorem 5, we obtain that

$$\mathbb{E}[h_{\mathcal{T}}(x_1)] = 1 + (1-\epsilon) \sum_{j=3}^{n} \frac{\epsilon^{j-2}}{1-\epsilon^{j-1}}.$$

and that for any $2 \leq i \leq n-1$,

$$\mathbb{E}[h_{\mathcal{T}}(x_i)] = (1-\epsilon) \left[\sum_{j=1}^{i-1} \frac{1}{1-\epsilon^j} + \sum_{j=0}^{n-1-i} \frac{\epsilon^j}{1-\epsilon^{j+1}} \right],$$

and finally that

$$\mathbb{E}[h_{\mathcal{T}}(x_n)] = (1-\epsilon) \sum_{j=1}^{n-1} \frac{1}{1-\epsilon^j}.$$

We recall that Δ_i is defined as $\Delta_i = \mathbb{E}[h_{\mathcal{T}}(x_{i+1})] - \mathbb{E}[h_{\mathcal{T}}(x_i)]$, for $i = 1, \dots, n-1$. From the above relations we obtain that

$$\Delta_1 = (1-\epsilon) \left[\frac{1}{1-\epsilon} - \frac{\epsilon^{n-2}}{1-\epsilon^{n-1}} \right] = 1 - (1-\epsilon) \frac{\epsilon^{n-2}}{1-\epsilon^{n-1}},$$

that for any $2 \leq i \leq n-2$,

$$\Delta_i = (1-\epsilon) \left[\frac{1}{1-\epsilon^i} - \frac{\epsilon^{n-1-i}}{1-\epsilon^{n-i}} \right],$$

and that

$$\Delta_{n-1} = (1-\epsilon) \left[\frac{1}{1-\epsilon^{n-1}} - \frac{1}{1-\epsilon} \right] = -\epsilon \frac{1-\epsilon^{n-2}}{1-\epsilon^{n-1}}.$$

Therefore $|\Delta_1 - 1| \leq \epsilon^{n-2}$ and $-\epsilon \leq \Delta_{n-1} \leq 0$ and for any $2 \leq i \leq n-2$,

$$|\Delta_i - 1| \leq \left| \frac{1-\epsilon}{1-\epsilon^i} - 1 \right| + \epsilon^{n-1-i} \leq 2\epsilon.$$

Using this, we conclude that $|\Delta_i - 1| \leq 2\epsilon$ for any $1 \leq i \leq n-2$ and that $-\epsilon \leq \Delta_{n-1} \leq 0$.

Proof of (18). Given some integer $3 < k < n-2$, the points in this configuration are such that $x_k = \frac{1}{2}$ and such that $\{x_1, \dots, x_{k-1}\}$ and $\{x_{k+1}, \dots, x_n\}$ extend over the intervals $[0, \epsilon]$ and $[1-\epsilon, 1]$ respectively. For the average height of x_k, using Theorem 5, we have

$$\begin{aligned}\mathbb{E}[h_{\mathcal{T}}(x_k)] &= 2 + \sum_{j=1}^{k-2} \frac{x_{j+1} - x_j}{x_k - x_j} + \sum_{j=k+2}^{n} \frac{x_j - x_{j-1}}{x_j - x_k} \\ &\leq 2 + \log\left(\frac{x_k - x_1}{x_k - x_{k-1}}\right) + \log\left(\frac{x_n - x_k}{x_{k+1} - x_k}\right),\end{aligned}$$

where the inequality follows from integral-series comparison with the function $x \mapsto 1/(x - x_k)$. Since $x_k - x_{k-1} = x_{k+1} - x_k = \frac{1}{2} - \epsilon$ and $x_n - x_{k+1} = \epsilon$ as well as $x_{k-1} - x_1 = \epsilon$, we obtain that

$$\mathbb{E}[h_{\mathcal{T}}(x_k)] \leq 2 + 2\log\left(1 + \frac{\epsilon}{\frac{1}{2} - \epsilon}\right) \leq 2 + 2\frac{\epsilon}{\frac{1}{2} - \epsilon} \leq 2 + 8\epsilon,$$

where the last inequality holds since $\epsilon < 1/4$. For $i = 1$, we have

$$\mathbb{E}[h_{\mathcal{T}}(x_1)] \geq 1 + \sum_{j=k}^{k+1} \frac{x_j - x_{j-1}}{x_j - x_1} = 1 + \frac{1/2 - \epsilon}{1/2} + \frac{1/2 - \epsilon}{1 - \epsilon} \geq \frac{5}{2} - 3\epsilon,$$

and for any $2 \leq i < k$, we have

$$\mathbb{E}[h_{\mathcal{T}}(x_i)] \geq 1 + \sum_{j=i+1}^{n} \frac{x_j - x_{j-1}}{x_j - x_i} \geq 2 + \frac{x_{k+1} - x_k}{x_{k+1} - x_i} \geq \frac{5}{2} - \epsilon.$$

Therefore $\mathbb{E}[h_{\mathcal{T}}(x_i)] \geq \frac{5}{2} - 3\epsilon$ for any $1 \leq i \leq k-1$ and by symetry this inequality holds for any $k+1 \leq i \leq n$.

References

Aggarwal, C. C. (2017). *Outlier analysis* (2nd edition). Cham: Springer.

Ahsanullah, M., Nevzorov, V. B., & Shakil, M. (2013). *An introduction to order statistics* (Vol. 3). Berlin: Springer.

Aragon, C., & Seidel, R. (1989). Randomized search trees. *30th Annual Symposium on Foundations of Computer Science.*

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data.* (3rd ed.). Chichester: John Wiley & Sons, Inc.

Chabchoub, Y., Togbe, M. U., Boly, A., & Chiky, R. (2022). An in-depth study and improvement of isolation forest. *IEEE Access*, *10*, 10219–10237.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, *41*(3), 1–58.

Devroye, L., Györfi, L., & Krzyżak, A. (1998). The hilbert kernel regression estimate. *Journal of Multivariate Analysis*, *65*(2), 209–227.

Devroye, L., & Krzyżak, A. (1999). On the hilbert kernel density estimate. *Statistics & Probability Letters*, *44*(3), 299–308.

Diestel, R. (2017). *Graph theory* (5th edition, Vol. 173). Berlin: Springer.

Drmota, M. (2009). *Random trees. An interplay between combinatorics and probability.* Wien: Springer.

Dvoretzky, A., Kiefer, J., & Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, *27*(3), 642–669.

Foorthuis, R. (2021). On the nature and types of anomalies: A review of deviations in data. *International Journal of Data Science and Analytics*, *12*(4), 297–331.

Hariri, S., Kind, M. C., & Brunner, R. J. (2021). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1479–1489.

Hibbard, T. N. (1962). Some combinatorial properties of certain trees with applications to searching and sorting. *Journal of the ACM*, *9*(1), 13–28.

Karczmarek, P., Kiersztyn, A., Pedrycz, W., & Al, E. (2020). K-means-based isolation forest. *Knowledge-Based Systems*, *195*, 105659.

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity*, *2*(1).

Lesouple, J., Baudoin, C., Spigai, M., & Tourneret, J.-Y. (2021). Generalized isolation forest for anomaly detection. *Pattern Recognition Letters*, *149*, 109–119.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2010). On detecting clustered anomalies using sciforest. In J. L. Balcázar, F. Bonchi, A. Gionis, & M. Sebag (Eds.), *Machine learning and knowledge discovery in databases* (pp. 274–290). Springer Berlin Heidelberg.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1–39.

Markou, M., & Singh, S. (2003a). Novelty detection: A review—part 1: Statistical approaches. *Signal Processing*, *83*(12), 2481–2497.

Markou, M., & Singh, S. (2003b). Novelty detection: A review—part 2: *Signal Processing*, *83*(12), 2499–2521.

Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, *18*(3).

Mensi, A., & Bicego, M. (2021). Enhanced anomaly scores for isolation forests. *Pattern Recognition*, *120*, 108115.

Mensi, A., Tax, D. M., & Bicego, M. (2023). Detecting outliers from pairwise proximities: Proximity isolation forests. *Pattern Recognition*, *138*, 109334.

Morales, F. A., Ramírez, J. M., & Ramos, E. A. (2022). A mathematical assessment of the isolation random forest method for anomaly detection in big data. *Mathematical Methods in the Applied Sciences*.

Mosler, K. (2013). Depth statistics. *Robustness and Complex Data Structures*, 17–34.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Preiss, B. R. (1999). *Data structures and algorithms with object-oriented design patterns in java*. Wiley.

Samariya, D., & Thakkar, A. (2023). A comprehensive survey of anomaly detection algorithms. *Annals of Data Science*, *10*(3), 829–850.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*(7), 1443–1471.

Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., & Platt, J. (1999). Support vector method for novelty detection. In S. Solla, T. Leen, & K. Müller (Eds.), *Advances in neural information processing systems* (Vol. 12). MIT Press.

Seidel, R., & Aragon, C. (1996). Randomized search trees. *Algorithmica*, *16*(4–5), 464–497.

Staerman, G., Mozharovskyi, P., Clémençon, S., & d'Alché-Buc, F. (2019, 17–19 Nov). Functional isolation forest. In W. S. Lee & T. Suzuki (Eds.), *Proceedings of the eleventh asian conference on machine learning* (pp. 332–347, Vol. 101). PMLR.

Wainwright, M. J. (2019). *High-dimensional statistics. A non-asymptotic viewpoint* (Vol. 48). Cambridge: Cambridge University Press.

Ziegler, G. M. (1995). *Lectures on polytopes* [Updated Seventh Printing of the First Edition]. Springer New York.