Minimum distance estimation with universal performance guarantees over statistical regular pavings

RAAZESH SAINUDIIN[†°] and GLORIA TENG[‡], [°]Department of Mathematics, Uppsala University, Uppsala, Sweden, and [‡]NEXT Academy, Kuala Lumpur, Malaysia.

We present a data-adaptive multivariate histogram estimator of an unknown density f based on n independent samples from it. Such data-dependent adaptive histograms are formalized as statistical regular pavings (SRPs). Regular pavings (RPs) are binary trees obtained by selectively bisecting a d-dimensional box in a recursive and regular manner. SRP augments RP by mutably caching the sufficient statistics of the data. Using a form of tree matrix collation with SRPs we obtain the minimum distance estimate (MDE) with universal performance guarantees over Yatracos classes and demonstrate the performance of the estimator with simulated data.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: —*Probabilistic algorithms (including Monte Carlo)*; *Statistical computing*; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Trees*; E.1 [**Data Structures**]: —*Trees*

General Terms: Algorithms, Design, Performance, Theory

Additional Key Words and Phrases: Rooted planar binary tree, Yatracos class, Tree matrix arithmetic

1. INTRODUCTION

Suppose our random variable X has an unknown density f on $\mathbb{R}^d$, then for all Borel sets $A \subseteq \mathbb{R}^d$,

$$\mu(A) := \Pr\{X \in A\} = \int_A f(x)dx \ .$$

Any density estimate $f_n(x) := f_n(x; X_1, X_2, \ldots, X_n) : \mathbb{R}^d \times \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ is a map from $\left(\mathbb{R}^d\right)^{n+1}$ to $\mathbb{R}$. The objective in density estimation is to estimate the unknown f from an independent and identically distributed (IID) sample $X_1, X_2, \ldots, X_n$ drawn from f. Density estimation is often the first step in many learning tasks, including, anomaly detection, classification, regression and clustering.

The quality of f_n is naturally measured by how well it performs the assigned task of computing the probabilities of sets under the total variation criterion:

$$\mathrm{TV}(f_n, f) = \sup_{A \in \mathcal{B}^d} \left| \int_A f_n - \int_A f \right| = \frac{1}{2} \int |f_n - f| \ .$$

The last equality above is due to Scheffé's identity and this equates the L_1 distance between f_n and f, in the absolute scale of $[0, 1]$, to the total variation distance between them.

A non-parametric density estimator is said to have *universal performance guarantees* if it is valid no matter what the underlying f happens to be [Devroye and Lugosi 2001, p. 1]. Histograms and kernel density estimators can approximate f in this universal sense in an asymptotic setting, i.e., as the number of data points n approaches infinity (the so-called *asymptotic consistency* of the estimator f_n). But for a fixed n, however large but finite, classical studies of the rate of convergence of f_n to f require additional assumptions on the smoothness class (to solve this so-called *smoothing problem*), such as $f \in L_2 \neq L_1$ or $f \in C^k$, the set of k-times differentiable functions, for some $k \geq 0$, as opposed to letting f simply belong to the set where densities exist, i.e., $f \in L_1$, and thereby violate the universality property.

Author's addresses: [†] Corresponding Author: Raazesh Sainudiin, [°]Department of Mathematics, Uppsala University, Box 480, 751 06 Uppsala, Sweden. [‡]NEXT Academy, Kuala Lumpur 60000, Malaysia.

For a concrete class of unknown densities of pressing interest in current applications involving periodic bursts of sample size $n \approxeq 10^7$, consider an $f \notin C^0$, where $f \ll \lambda$, but its inverse image at 0 has finitely many distinct Lebesgue-measurable full compact sets (say, h distinct "clean holes"), within a box $\check{\boldsymbol{x}} := [-M, M]^d$ for a large enough M, such that, $\mu(\check{\boldsymbol{x}}^c) < \xi$, where $\check{\boldsymbol{x}}^c := \mathbb{R}^d \setminus \check{\boldsymbol{x}}$, for any given but fixed $\xi > 0$. Thus, $f^{[-1]}(0) := \{x : f(x) = 0\} = \cup_{i=1}^{h} H_i$, where each $H_i \subset \check{\boldsymbol{x}}$ and $H_i \cap H_j = \emptyset$, for $i \neq j$. Such an f is in L_1, given that it is a density, and crucially, an estimator f_n with universal performance guarantees will estimate $f \in L_1$ and *not* merely assure asymptotic L_1 consistency while mathematically compromising to f actually *not* being in L_1 in order to solve the smoothing problem that is also required to obtain f_n for a given $n < \infty$.

Universal performance guarantee is provided by the *minimum distance estimate* (MDE) due to [Devroye and Lugosi 2001; 2004]. Their fundamentally combinatorial approach combined ideas from [Yatracos 1985; 1988] on minimum distance methods and from Vapnik and Chervonenkis [Vapnik and Chervonenkis 1971] on uniform convergence of empirical probability measures over classes of sets.

Tree based partitioning strategies are particularly suited to large sample sizes and we focus on tree based histograms here using statistical regular pavings. The particular class of MDEs studied in [Devroye and Lugosi 2001; 2004] were limited to kernel estimates and histograms under simpler partitioning rules. Inspired by this, here we develop an MDE over statistical regular pavings to produce nonparametric data-adaptive density estimates in d dimensions with universal performance guarantees.

Our approach exploits a recursive arithmetic using nodes imbued with recursively computable statistics and a specialized collator structure to compute the supremal deviation of the held-out empirical measure over the Yatracos class of the candidate densities. Although a more efficient algorithm (up to pre-processing the L_1 distances for each pair of densities) is characterized in [Mahalanabis and Stefankovic 2008], we are not aware of any implementations of the MDE using data-adaptive multivariate histograms for bursts of data (with $n \approxeq 10^7$ in dimensions up to 6 for instance in a non-distributed computational setting over one commodity machine). To the best of our knowledge, the accompanying code of this paper in `mrs2` [Sainudiin et al. 2018] is the only publicly available implementation of such an MDE estimator.

2. REGULAR PAVINGS AND HISTOGRAMS

Let $\boldsymbol{x} := [\underline{x}, \overline{x}]$ be a compact real interval with lower bound $\underline{x}$ and upper bound $\overline{x}$, where $\underline{x} \leq \overline{x}$. Let the space of such intervals be $\mathbb{IR}$. The width of an interval $\boldsymbol{x}$ is $\mathrm{wid}\,(\boldsymbol{x}) := \overline{x} - \underline{x}$. The midpoint is $\mathrm{mid}\,(\boldsymbol{x}) := (\underline{x} + \overline{x})/2$. A box of dimension d with coordinates in $\Delta := \{1, 2, \ldots, d\}$ is an interval vector with ι as the first coordinate of maximum width:

$$\boldsymbol{x} := [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_d, \overline{x}_d] =: \underset{j \in \Delta}{\otimes} [\underline{x}_j, \overline{x}_j], \quad \iota := \min\left(\underset{i}{\mathrm{argmax}}(\mathrm{wid}\,(\boldsymbol{x}_i))\right) \ .$$

The set of all such boxes is $\mathbb{IR}^d$, i.e., the set of all interval real vectors in dimension d. A *bisection* or *split* of $\boldsymbol{x}$ perpendicularly at the mid-point along this first widest coordinate ι gives the left and right child boxes of $\boldsymbol{x}$

$$\boldsymbol{x}_{\mathsf{L}} := [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_\iota, \mathrm{mid}\,(\boldsymbol{x}_\iota)) \times [\underline{x}_{\iota+1}, \overline{x}_{\iota+1}] \times \ldots \times [\underline{x}_d, \overline{x}_d] \ ,$$

$$\boldsymbol{x}_{\mathsf{R}} := [\underline{x}_1, \overline{x}_1] \times \ldots \times [\mathrm{mid}\,(\boldsymbol{x}_\iota), \overline{x}_\iota] \times [\underline{x}_{\iota+1}, \overline{x}_{\iota+1}] \times \ldots \times [\underline{x}_d, \overline{x}_d] \ .$$

Such a bisection is said to be *regular*. Note that this bisection gives the left child box a half-open interval $[\underline{x}_\iota, \mathrm{mid}\,(\boldsymbol{x}_\iota))$ on coordinate ι so that the intersection of the left and right child boxes is empty. A recursive sequence of selective regular bisections of boxes, with possibly open boundaries, along the first widest coordinate, starting

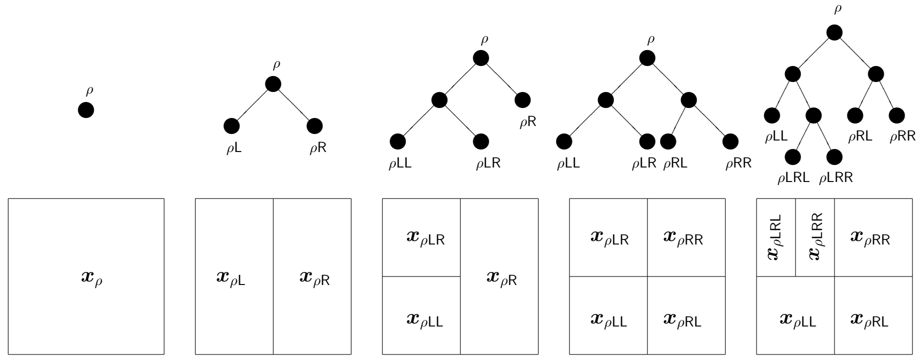

ρ
ρ
ρL
ρR
ρ
ρLL
ρLR
ρR
ρ
ρLL
ρLR
ρRL
ρRR
ρ
ρLL
ρRL
ρRR
ρLRL
ρLRR
$\boldsymbol{x}_\rho$
$\boldsymbol{x}_{\rho\mathsf{L}}$
$\boldsymbol{x}_{\rho\mathsf{R}}$
$\boldsymbol{x}_{\rho\mathsf{LR}}$
$\boldsymbol{x}_{\rho\mathsf{LL}}$
$\boldsymbol{x}_{\rho\mathsf{R}}$
$\boldsymbol{x}_{\rho\mathsf{LR}}$
$\boldsymbol{x}_{\rho\mathsf{RR}}$
$\boldsymbol{x}_{\rho\mathsf{LL}}$
$\boldsymbol{x}_{\rho\mathsf{RL}}$
$\boldsymbol{x}_{\rho\mathsf{LRL}}$
$\boldsymbol{x}_{\rho\mathsf{LRR}}$
$\boldsymbol{x}_{\rho\mathsf{RR}}$
$\boldsymbol{x}_{\rho\mathsf{LL}}$
$\boldsymbol{x}_{\rho\mathsf{RL}}$

Fig. 1. A sequence of selective bisections of boxes (nodes) along the first widest coordinate, starting from the root box (root node) in two dimensions, produces an RP.

from the root box $\boldsymbol{x}_\rho$ in $\mathbb{IR}^d$ is known as a *regular paving* [Kieffer et al. 2001] or n-tree [Samet 1990] of $\boldsymbol{x}_\rho$. A regular paving of $\boldsymbol{x}_\rho$ can also be seen as a binary tree formed by recursively bisecting the box $\boldsymbol{x}_\rho$ at the root node. Each node in the binary tree has either no children or two children. These trees are known as plane binary trees in enumerative combinatorics [Stanley 1999, Ex. 6.19(d), p. 220] and as finite, rooted binary trees (frb-trees) in geometric group theory [Meier 2008, Chap. 10]. The relationship of trees, labels and partitions is illustrated in Figure 1 via a sequence of bisections of a square (2-dimensional) root box by always bisecting on the *first* widest coordinate.

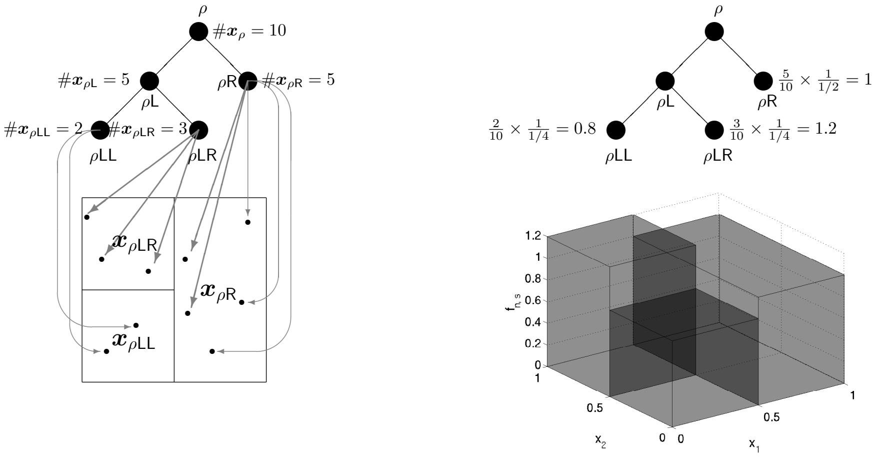
Let $\mathbb{N} := \{1, 2, \ldots\}$ be the set of natural numbers. Let the j-th interval of a box $\boldsymbol{x}_{\rho\mathsf{v}}$ be $[\underline{x}_{\rho\mathsf{v},j}, \overline{x}_{\rho\mathsf{v},j}]$, the volume of a d-dimensional box $\boldsymbol{x}_{\rho\mathsf{v}}$ be $\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}}) = \prod_{j=1}^{d}(\overline{x}_{\rho\mathsf{v},j} - \underline{x}_{\rho\mathsf{v},j})$, the set of all nodes of an RP be $\mathbb{V} := \rho \cup \{\rho\{\mathsf{L},\mathsf{R}\}^j : j \in \mathbb{N}\}$, the set of all leaf nodes be $\mathbb{L}$ and the set of internal nodes or splits be $\breve{\mathbb{V}}(s) := \mathbb{V}(s) \setminus \mathbb{L}(s)$. The set of leaf boxes of a regular paving s with root box $\boldsymbol{x}_\rho$ is denoted by $\boldsymbol{x}_{\mathbb{L}(s)}$ and it specifies a partition of the root box $\boldsymbol{x}_\rho$. Let $\mathbb{S}_k$ be the set of all regular pavings with root box $\boldsymbol{x}_\rho$ made of k splits. Note that the number of leaf nodes $m = |\mathbb{L}(s)| = k + 1$ if $s \in \mathbb{S}_k$. The number of distinct binary trees with k splits is equal to the Catalan number C_k.

$$C_k = \frac{1}{k+1}\binom{2k}{k} = \frac{(2k)!}{(k+1)!(k!)} \quad . \tag{1}$$

For $i, j \in \mathbb{Z}_+$, where $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$ and $i \leq j$, let $\mathbb{S}_{i:j} := \cup_{k=i}^{j}\mathbb{S}_k$ be the set of regular pavings with k splits where $k \in \{i, i+1, \ldots, j\}$. Let the set of all regular pavings be $\mathbb{S}_{0:\infty} := \lim_{j\to\infty}\mathbb{S}_{0:j}$.

A *statistical regular paving* (SRP) denoted by s is an extension of the RP structure that is able to act as a partitioned 'container' and responsive summarizer for multivariate data. An SRP can be used to create a histogram of a data set. A recursively computable statistic [Fisher 1925; Gray and Moore 2003] that an SRP node $\rho\mathsf{v}$ caches is $\#\boldsymbol{x}_{\rho\mathsf{v}}$, the count of the number of data points that fell into $\boldsymbol{x}_{\rho\mathsf{v}}$. A leaf node $\rho\mathsf{v}$ with $\#\boldsymbol{x}_{\rho\mathsf{v}} > 0$ is a non-empty leaf node. The set of non-empty leaves of an SRP s is $\mathbb{L}^+(s) := \{\rho\mathsf{v} \in \mathbb{L}(s) : \#\boldsymbol{x}_{\rho\mathsf{v}} > 0\} \subseteq \mathbb{L}(s)$.

Figure 2 depicts a small SRP s with root box $\boldsymbol{x}_\rho \in \mathbb{IR}^2$. The number of sample data points in the root box $\boldsymbol{x}_\rho$ is 10. Figure 2(a) shows the tree, including the count associated with each node in the tree and the partition of the root box represented by the leaf boxes of this tree, with the sample data points superimposed on the boxes. Figure 2(b) shows how the density estimate is computed from the count and the volume of leaf boxes to obtain the density estimate $f_{n,s}$ as an SRP histogram.


ρ
$\#\boldsymbol{x}_\rho = 10$
$\#\boldsymbol{x}_{\rho\mathsf{L}} = 5$
$\rho\mathsf{L}$
$\rho\mathsf{R}$
$\#\boldsymbol{x}_{\rho\mathsf{R}} = 5$
$\#\boldsymbol{x}_{\rho\mathsf{LL}} = 2$
$\#\boldsymbol{x}_{\rho\mathsf{LR}} = 3$
$\rho\mathsf{LL}$
$\rho\mathsf{LR}$
$\boldsymbol{x}_{\rho\mathsf{LR}}$
$\boldsymbol{x}_{\rho\mathsf{R}}$
$\boldsymbol{x}_{\rho\mathsf{LL}}$
ρ
$\frac{5}{10} \times \frac{1}{1/2} = 1$
$\rho\mathsf{L}$
$\rho\mathsf{R}$
$\frac{2}{10} \times \frac{1}{1/4} = 0.8$
$\frac{3}{10} \times \frac{1}{1/4} = 1.2$
$\rho\mathsf{LL}$
$\rho\mathsf{LR}$
$f_{n,s}$
1.2
1
0.8
0.6
0.4
0.2
0
1
0.5
0
x_2
0
0.5
1
x_1

(a) An SRP tree and its constituents.

(b) An SRP histogram $f_{n,s}$.

Fig. 2. An SRP and its corresponding histogram.

An SRP histogram is obtained from n data points that fell into $\boldsymbol{x}_\rho$ of SRP s as follows:

$$f_{n,s}(x) = f_n(x) = \sum_{\rho\mathsf{v} \in \mathbb{L}(s)} \frac{\mathbb{1}_{\boldsymbol{x}_{\rho\mathsf{v}}}(x)}{n} \left(\frac{\#\boldsymbol{x}_{\rho\mathsf{v}}}{\mathrm{vol}\left(\boldsymbol{x}_{\rho\mathsf{v}}\right)} \right) \quad . \tag{2}$$

It is the maximum likelihood estimator over the class of simple (piecewise-constant) functions given the partition $\boldsymbol{x}_{\mathbb{L}(s)}$ of the root box of s. We suppress subscripting the histogram by the SRP s for notational convenience. SRP histograms have some similarities to dyadic histograms (for eg. [Klemelä 2009, chap. 18], [Lu et al. 2013]). Both are binary tree-based and partition so that a box may only be bisected at the midpoint of one of its coordinates, but the RP structure restricts partitioning further by only bisecting a box on its first widest coordinate in order to make $\mathbb{S}_{0:\infty}$ closed under addition and scalar multiplication and thereby allowing for computationally efficient computer arithmetic over a dense set of simple functions (see [Harlow et al. 2012] for statistical applications of this arithmetic). Crucially, when data bursts have large sample sizes, this restrictive partitioning does not affect the L_1 errors when compared to a computationally more expensive Bayes estimator (see Sec. 4).

A statistically equivalent block (SEB) partition of a sample space is some partitioning scheme that results in equal numbers of data points in each element (block) of the partition [Tukey 1947]. The output of SEBTreeMC$(s, \overline{\#}, \overline{m})$ of Algorithm 1 is $[s(0), s(1), \ldots, s(T)]$, a sequence of SRP states visited by a sample path of the Markov chain $\{S(t)\}_{t \in \mathbb{Z}_+}$ on $\mathbb{S}_{0:\overline{m}-1}$, such that, $\mathbb{L}^\triangledown(s(T)) = \emptyset$, or $\#(\rho\mathsf{v}) \leq \overline{\#}\ \forall \rho\mathsf{v} \in \mathbb{L}^\triangledown(s(T))$, or $|\mathbb{L}(s(T))| = \overline{m}$ and T is a corresponding random stopping time. As the initial state $S(t = 0)$ is the root $s \in \mathbb{S}_0$, the Markov chain $\{S(t)\}_{t \in \mathbb{Z}_+}$ on $\mathbb{S}_{0:\overline{m}-1}$ satisfies $S(t) \in \mathbb{S}_t$ for each $t \in \mathbb{Z}_+$, i.e., the state at time t has $t+1$ leaves or t splits. The operation may only be considered to be successful if $|\mathbb{L}(s)| \leq \overline{m}$ and $\#\boldsymbol{x}_{\rho\mathsf{v}} \leq \overline{\#}\ \forall \rho\mathsf{v} \in \mathbb{L}^\triangledown(s)$. Therefore, the sequence of SRP histogram states visited by SEBTreeMC that successfully terminates at

ALGORITHM 1: SEBTreeMC$(s, \overline{\#}, \overline{m})$

```
input      : $s$, initial SRP with root node $\rho$,
             $x = (x_1, x_2, \ldots, x_n)$, a data burst of size $n$,
             $\# : \mathbb{L}^{\triangledown}(s) \rightarrow \mathbb{R}$, a priority function of counts,
             $\overline{\#}$, maximum value of $\#(\rho\mathsf{v}) \in \mathbb{L}^{\triangledown}(s)$ for any splittable leaf node in the final SRP,
             $\overline{m}$, maximum number of leaves in the final SRP.
output     : a sequence of SRP states $[s(0), s(1), \ldots, s(T)]$ such that $\mathbb{L}^{\triangledown}(s(T)) = \emptyset$ or $\#(\rho\mathsf{v}) \leq \overline{\#}$
             $\forall \rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s(T))$ or $|\mathbb{L}(s(T))| = \overline{m}$ .
initialize: $\boldsymbol{x}_\rho \leftsquigarrow x$, make $\boldsymbol{x}_\rho$ such that $\cup_i^n x_i \subset \boldsymbol{x}_\rho$ if $\nexists$ domain knowledge or historical data,
             $s \leftsquigarrow \boldsymbol{x}_\rho$, specify the root box of $s$,
             s $\leftarrow [s]$
while $\mathbb{L}^{\triangledown}(s) \neq \emptyset$ & $|\mathbb{L}(s)| < \overline{m}$ & $\max_{\rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s)} \#(\rho\mathsf{v}) > \overline{\#}$ do
    $\rho\mathsf{v} \leftarrow$ random_sample$\left(\underset{\rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s)}{\operatorname{argmax}} \#(\rho\mathsf{v})\right)$   // sample uniformly from nodes with largest #
    $s \leftarrow s$ with node $\rho\mathsf{v}$ split                 // split the sampled node and update $s$
    s.append($s$)                 // append the new SRP state with an additional split
end
```

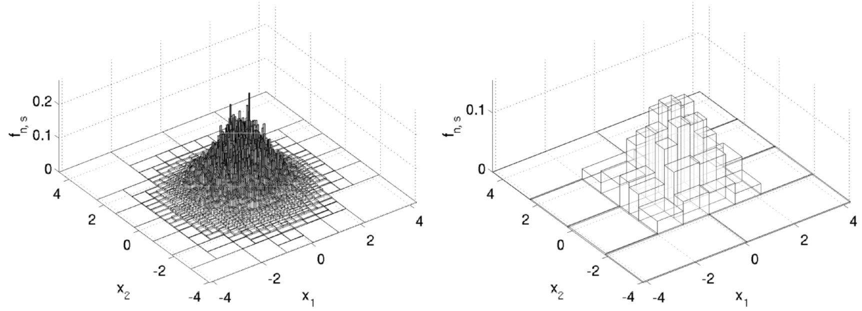


Fig. 3. Two histogram density estimates for the standard bivariate Gaussian density. The left figure shows a histogram with 1485 leaf nodes where $\overline{\#} = 50$ and the histogram on the right has $\overline{\#} = 1500$ resulting in 104 leaf nodes.

stopping time T will have the terminal histogram with at most $\overline{\#}$ many of the n data points in each of its leaf nodes and with at most $\overline{m}$ many leaf nodes.

Intuitively, SEBTreeMC$(s, \overline{\#}, \overline{m})$ prioritizes the splitting of leaf nodes with the largest numbers of data points associated with them. As we will see in Theorem 3.1, the L_1 consistency of SEBTreeMC requires that $\overline{m}$ must grow sublinearly (i.e. $\overline{m}/n \rightarrow 0$ as $n \rightarrow \infty$) while the volume of leaf boxes shrink such that a combinatorial complexity measure of the partitions in the support of the SEBTreeMC grows sub-exponentially. Figure 3 shows two different SRP histograms constructed using two different values of $\overline{\#}$ for the same dataset of $n = 10^5$ points simulated under the standard bivariate Gaussian density. A small $\overline{\#}$ produces a histogram that is under-smoothed with unnecessary spikes (Fig. 3 left) while the other histogram with a larger $\overline{\#}$ is over-smoothed (Fig. 3 right). We will obtain the minimum distance estimate from the SRP histograms visited by the SEBTreeMC in Theorem 3.3.

3. MINIMUM DISTANCE ESTIMATION USING STATISTICAL REGULAR PAVINGS

We show that the SRP density estimate from the `SEBTreeMC`-based partitioning scheme is asymptotically L_1-consistent as $n \to \infty$ provided that $\overline{\#}$, the maximum sample size in any leaf box in the partition, and $\overline{m}$, the maximum number of leaf boxes in the partition, grow with the sample size n at appropriate rates. This is done by proving the three conditions in Theorem 1 of [Lugosi and Nobel 1996]. We will need to show that as the number of sample points increases linearly, the following conditions are met:

(1) the number of leaf boxes grows sub-linearly;
(2) the partition grows sub-exponentially in terms of a combinatorial complexity measure;
(3) and the volume of the leaf boxes in the partition are shrinking.

Let $\{S_n(i)\}_{i=0}^{\dot{I}}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain of algorithm `SEBTreeMC`. The Markov chain terminates at some state $\dot{s}$ with partition $\mathbb{L}(\dot{s})$. Associated with the Markov chain is a fixed collection of partitions

$$\mathcal{L}_n := \left\{ \mathbb{L}(\dot{s}) : \dot{s} \in \mathbb{S}_{0:\infty}, \Pr\{S(\dot{I}) = \dot{s}\} > 0 \right\}$$

and the size of the largest partition $\mathbb{L}(\dot{s})$ in $\mathcal{L}_n$ is given by

$$m(\mathcal{L}_n) := \sup_{\mathbb{L}(\dot{s}) \in \mathcal{L}_n} |\mathbb{L}(\dot{s})| \leq \overline{m}$$

such that $\mathcal{L}_n \subseteq \{\mathbb{L}(s) : s \in \mathbb{S}_{0:\overline{m}-1}\}$.

Given n fixed points $\{x_1, \ldots, x_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\Pi\left(\mathcal{L}_n, \{x_1, \ldots, x_n\}\right)$ be the number of distinct partitions of the finite set $\{x_1, \ldots, x_n\}$ that are induced by partitions $\mathbb{L}(\dot{s}) \in \mathcal{L}_n$:

$$\Pi(\mathcal{L}_n, \{x_1, \ldots, x_n\}) := |\{\{\boldsymbol{x}_{\rho v} \cap \{x_1, \ldots, x_n\} : \boldsymbol{x}_{\rho v} \in \mathbb{L}(\dot{s})\} \ : \mathbb{L}(\dot{s}) \in \mathcal{L}_n\}| \ .$$

For any fixed set of n points, the growth function of $\mathcal{L}_n$ is then

$$\Pi^*(\mathcal{L}_n, \{x_1, \ldots, x_n\}) = \max_{\{x_1, \ldots, x_n\} \in (\mathbb{R}^d)^n} \Pi(\mathcal{L}_n, \{x_1, \ldots, x_n\}) \ .$$

Let $A \subseteq \mathbb{R}^d$. Then the diameter of A is the maximum Euclidean distance between any two points of A, i.e., $\mathrm{diam}(A) := \sup_{x,y \in A} \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$. Thus, for a box $\boldsymbol{x} = [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_d, \overline{x}_d]$, $\mathrm{diam}(\boldsymbol{x}) = \sqrt{\sum_{i=1}^d (\overline{x}_i - \underline{x}_i)^2}$.

THEOREM 3.1 (L_1-CONSISTENCY). *Let $X_1, X_2, \ldots$ be independent and identical random vectors in $\mathbb{R}^d$ whose common distribution μ has a non-atomic density f, i.e., $\mu \ll \lambda$. Let $\{S_n(i)\}_{i=0}^{\dot{I}}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain formed using* `SEBTreeMC` *(Algorithm 1) with terminal state $\dot{s}$ and histogram estimate $f_{n,\dot{s}}$ over the collection of partitions $\mathcal{L}_n$. As $n \to \infty$, if $\overline{\#} \to \infty$, $\overline{\#}/n \to 0$, $\overline{m} \geq n/\overline{\#}$, and $\overline{m}/n \to 0$ then the density estimate $f_{n,\dot{s}}$ is asymptotically consistent in L_1, i.e.,*

$$\int |f(x) - f_{n,\dot{s}}(x)| dx \to 0 \textit{ with probability } 1.$$

PROOF. We will assume that $\overline{\#} \to \infty$, $\overline{\#}/n \to 0$, $\overline{m} \geq n/\overline{\#}$, and $\overline{m}/n \to 0$, as $n \to \infty$, and show that the three conditions:

(a) $n^{-1} m(\mathcal{L}_n) \to 0,$
(b) $n^{-1} \log \Pi_n^*(\mathcal{L}_n) \to 0,$ and
(c) $\mu(x : \mathrm{diam}(\boldsymbol{x}(x)) > \gamma) \to 0$ with probability 1 for every $\gamma > 0$,

are satisfied. Then by Theorem 1 of Lugosi and Nobel (1996) our density estimate $f_{n,\dot{s}}$ is asymptotically consistent in L_1.

Condition (a) is satisfied by the assumption that $\overline{m}/n \to 0$ since $m(\mathcal{L}_n) \leq \overline{m}$.

The largest number of distinct partitions of any n point subset of $\mathbb{R}^d$ that are induced by the partitions in $\mathcal{L}_n$ is upper bounded by the size of the collection of partitions $\mathcal{L}_n \subseteq \mathbb{S}_{0:\overline{m}-1}$, i.e.,

$$\Pi_n^*(\mathcal{L}_n) \leq |\mathcal{L}_n| \leq \sum_{k=0}^{\overline{m}-1} C_k$$

where k is the number of splits.

The growth function is thus bounded by the total number of partitions with 0 to $\overline{m}-1$ splits, i.e., the $(\overline{m}-1)$-th partial sum of the Catalan numbers. The partial sum can be asymptotically equivalent to ([Mattarei 2010]):

$$\sum_{k=0}^{\overline{m}-1} C_k \sim \frac{4^{\overline{m}}}{\left(3(\overline{m}-1)\sqrt{\pi(\overline{m}-1)}\right)} \quad \text{as } \overline{m} \to \infty \ .$$

Taking logs and dividing by n on both sides of the above two equations, and using the assumption that $\overline{m}/n \to 0$ as $n \to \infty$, we can see that condition (b) is satisfied:

$$\log \Pi_n^*(L_n)/n \ \leq \log(|\mathcal{L}_n|)/n \to \tfrac{1}{n}\left(\overline{m}\log 4 - \tfrac{3}{2}\log(\overline{m}-1) - \log 3\sqrt{\pi}\right) \to 0.$$

We now prove the final condition. Fix $\gamma, \xi > 0$. There exists a box $\check{\boldsymbol{x}} = [-M, M]^d$ for a large enough M, such that, $\mu(\check{\boldsymbol{x}}^c) < \xi$, where $\check{\boldsymbol{x}}^c := \mathbb{R}^d \setminus [-M, M]^d$. Consequently,

$$\mu(\{x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma\}) \leq \xi + \mu(\{x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma\} \cap \check{\boldsymbol{x}}).$$

Using 2^{di} hypercubes of equal volume $(2M)^d/2^{di}, i = \left\lceil \log_2\left(2M\sqrt{d}/\gamma\right)\right\rceil$ with side length $2M/2^i$ and diameter $\sqrt{d(\frac{2M}{2^i})^2}$, we can have at most $m_\gamma < 2^{di}$ boxes in $\check{x}$ that have diameter greater than γ. By choosing i large enough we can upper bound m_γ by $(2M\sqrt{d}/\gamma)^d$, a quantity that is independent of n, such that

$$\begin{aligned}
\mu(x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma) \ &\leq \ \xi + \mu\left(\{x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma\} \cap \check{\boldsymbol{x}}\right) \\
&\leq \ \xi + m_\gamma\left(\max_{\boldsymbol{x} \in \mathbb{L}(\dot{s})} \mu(\boldsymbol{x})\right) \\
&\leq \ \xi + m_\gamma\left(\max_{\boldsymbol{x} \in \mathbb{L}(\dot{s})} \mu_n(\boldsymbol{x}) + \max_{\boldsymbol{x} \in \mathbb{L}(\dot{s})} |\mu(\boldsymbol{x}) - \mu_n(\boldsymbol{x})|\right),\ \mu_n(\boldsymbol{x}) := \frac{\#(\boldsymbol{x})}{n} \\
&\leq \ \xi + m_\gamma\left(\frac{\overline{\#}}{n} + \sup_{\boldsymbol{x} \in \mathrm{I\!R}^d} |\mu(\boldsymbol{x}) - \mu_n(\boldsymbol{x})|\right).
\end{aligned}$$

The first term in the parenthesis converges to zero since $\overline{\#}/n \to 0$ by assumption. For $\epsilon > 0$ and $n > 4d$, the second term goes to zero by applying the Vapnik-Chervonenkis (VC) theorem to boxes in $\mathrm{I\!R}^d$ with VC dimension $2d$ and shatter coefficient $S(\mathrm{I\!R}^d, n) \leq (en/2d)^{2d}$ [Devroye et al. 1996, Thms. 12.5, 13.3 and p. 220], i.e.,

$$\Pr\left\{\sup_{\boldsymbol{x} \in \mathrm{I\!R}^d} |\mu_n(\boldsymbol{x}) - \mu(\boldsymbol{x})| > \epsilon\right\} \leq 8 \cdot (en/2d)^{2d} \cdot e^{-n\epsilon^2/32} \ .$$

For any $\epsilon > 0$ and finite d, the right-hand-side of the above inequality can be made arbitrarily small for n large enough. This convergence in probability is equivalent to the following almost sure convergence by the bounded difference inequality:

$$\lim_{n\to\infty} \sup_{\boldsymbol{x}\in\mathbb{R}^d} |\mu_n(\boldsymbol{x}) - \mu(\boldsymbol{x})| = 0 \quad \text{w.p. } 1 \enspace .$$

Thus for any $\gamma, \xi > 0$,

$$\lim_{n\to\infty} \mu(\{x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma\}) \leq \xi \quad \text{w.p. } 1 \enspace .$$

Therefore, condition (c) is satisfied and this completes the proof. □

Let Θ index a set of finitely many density estimates: $\{f_{n,\theta} : \theta \in \Theta\}$, such that $\int f_{n,\theta} = 1$ for each $\theta \in \Theta$. We can index the SRP trees by $\{s_\theta : \theta \in \Theta\}$, where θ is the sequence of leaf node depths that uniquely identifies the SRP tree, and denote the density estimate corresponding to s_θ by f_{n,s_θ} or simply by $f_{n,\theta}$. Now, consider the asymptotically consistent path taken by the Markov chain of `SEBTreeMC`. For a fixed sample size n, let $\{s_\theta : \theta \in \Theta\}$ be an ordered subset of states visited by the Markov chain, with $s_\theta \prec s_\vartheta$ if s_ϑ is a refinement of s_θ, i.e. if s_θ is visited before s_ϑ. The goal is to select the optimal estimate from $|\Theta|$ many candidates.

When our candidate set of densities are additive like the histograms, we can use the hold-out method proposed by Devroye and Lugosi [2001, Sec. 10.1] for minimum distance estimation as follows. Let $0 < \varphi < 1/2$. Given n data points, use $n - \varphi n$ points as the training set and the remaining φn points as the validation set (by φn we mean $\lfloor \varphi n \rfloor$). Denote the set of training data by $\mathcal{T} := \{x_1, \ldots, x_{n-\varphi n}\}$ and the set of validation data by $\mathcal{V} := \{x_{n-\varphi n+1}, \ldots, x_n\} = \{y_1, \ldots, y_{\varphi n}\}$. For an ordered pair $(\theta, \vartheta) \in \Theta^2$, with $\theta \neq \vartheta$, the set:

$$A_{\theta,\vartheta} := A\left(f_{n-\varphi n,\theta}, f_{n-\varphi n,\vartheta}\right) := \{x : f_{n-\varphi n,\theta}(x) > f_{n-\varphi n,\vartheta}(x)\}$$

is known as a *Scheffé set*. The *Yatracos class* [Yatracos 1985] is the collection of all such Scheffé sets over Θ:

$$\mathcal{A}_\Theta = \left\{\{x : f_{n-\varphi n,\theta}(x) > f_{n-\varphi n,\vartheta}(x)\} : (\theta,\vartheta) \in \Theta^2, \theta \neq \vartheta\right\} \enspace .$$

Let $\mu_{\varphi n}$ be the empirical measure of the validation set $\mathcal{V}$. Then the *minimum distance estimate* or MDE $f_{n-\varphi n,\theta^*}$ is the density estimate $f_{n-\varphi n,\theta}$ constructed from the training set $\mathcal{T}$ with the smallest index θ^* that minimizes:

$$\Delta_\theta = \sup_{A\in\mathcal{A}_\Theta} \left| \int_A f_{n-\varphi n,\theta}(A) - \mu_{\varphi n}(A) \right| \enspace . \tag{3}$$

Thus, the MDE $f_{n-\varphi n,\theta^*}$ minimizes the supremal absolute deviation from the held-out empirical measure $\mu_{\varphi n}$ over the Yatracos class $\mathcal{A}_\Theta$.

The SRP is adapted for MDE to mutably cache the counts for training and validation data separately and the $n - \varphi n$ training data points in $\mathcal{T}$ and the φn validation data points in $\mathcal{V}$ are accessible from any leaf node ρv of the SRP via pointers to $x_i \in \mathcal{T}$ and $y_i \in \mathcal{V}$, respectively. The training data drive the Markov chain `SEBTreeMC`$(s, \overline{\#}, \overline{m})$ to produce a sequence of SRP states: $s_{\theta_1}, s_{\theta_2}, \ldots$ that are further selected to build the candidate set of adaptive histogram density estimates given by $\{f_{n-\varphi n,\theta_i} : \theta_i \in \Theta\}$. For each $\theta_i \in \Theta$, the validation data is allowed to flow through s_{θ_i} and drop into the leaf boxes of s_{θ_i}. A graphical representation of an SRP with training counter $\#\boldsymbol{x}_{\rho v}$ and validation counter $\check{\#}\boldsymbol{x}_{\rho v}$ is shown in Figure 4. Computing the MDE objective Δ_{θ_i} in (3) requires the histogram estimate $f_{n-\varphi n}(\rho v) = \#\boldsymbol{x}_{\rho v}/n\lambda(\boldsymbol{x}_{\rho v})$ and the empirical measure of the validation data $\mu_{\varphi n}(\boldsymbol{x}_{\rho v}) = \check{\#}\boldsymbol{x}_{\rho v}/\varphi n$ at any node ρv. These can be readily obtained from $\#\boldsymbol{x}_{\rho v}$ and $\check{\#}\boldsymbol{x}_{\rho v}$.

Our approach to obtaining the MDE $f_{n-\varphi n,\theta^*}$ with optimal SRP s_{θ^*} exploits the partition refinement order in $\{s_\theta : \theta \in \Theta\}$, a subset of states along the path taken by

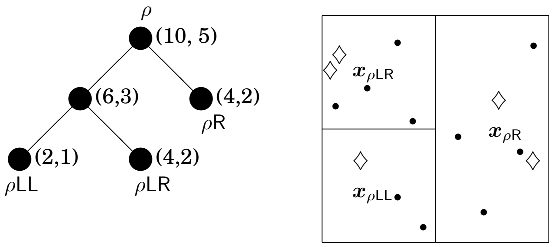

ρ
(10, 5)
(6,3)
(4,2)
ρR
(2,1)
(4,2)
ρLL
ρLR
$\boldsymbol{x}_{\rho\mathsf{LR}}$
$\boldsymbol{x}_{\rho\mathsf{R}}$
$\boldsymbol{x}_{\rho\mathsf{LL}}$

Fig. 4. An SRP s with training (•) and validation data (◇) and their respective sample counts ($\#\boldsymbol{x}_{\rho v}$, $\breve{\#}\boldsymbol{x}_{\rho v}$) that are updated recursively as data falls through the nodes of s.

the SEBTreeMC. Using nodes imbued with recursively computable statistics for both training and validation data, and a specialized collation according to SRPCollate (Algorithm 3) over SRPs, we compute the objective Δ_θ in (3) using GetDelta (Algorithm 2) via a dynamically grown Yatracos Matrix with pointers to all Scheffé sets constituting the Yatracos class according to GetYatracos (Algorithm 4). We briefly outline the core ideas in these three algorithms next (see Appendix for their pseudocode and mrs2 [Sainudiin et al. 2018] for details).

In the MDE procedure, pairwise comparisons of the heights of the candidate density estimates $f_{n-\varphi n,\theta}$ and $f_{n-\varphi n,\vartheta}$ are needed to get the Scheffé sets that make up the Yatracos class. An efficient way to approach this is to collate the SRPs corresponding to the density estimates onto a *collator regular paving* (CRP) where the space of CRP trees is also $\mathbb{S}_{0:\infty}$. Consider now two SRPs s_θ and s_ϑ for which the corresponding histogram estimates $f_{n,\theta}$ and $f_{n,\vartheta}$ are computed. Both SRPs s_θ and s_ϑ have the same root box $\boldsymbol{x}_\rho$. By collating the two SRPs we get a CRP c with the same root box and the tree obtained from a union of s_θ and s_ϑ. Unlike the union operation over RPs ([Harlow et al. 2012, Algorithm 1]), each node ρv of the SRP collator c stores $f_{n,\theta}$ and $f_{n,\vartheta}$ as a vector $\boldsymbol{f}_{n,c}(\rho v) := (f_{n,\theta}(\rho v), f_{n,\vartheta}(\rho v))$. The empirical measure of the validation data $\mu_{\varphi n}(\boldsymbol{x}_{\rho v})$ will also be stored at each node ρv and can be easily accessed via pointers. Figure 5 shows how CRP c can collate two SRPs s_θ and s_ϑ using SRPCollate.

We now use Theorem 10.1 of [Devroye and Lugosi 2001, p. 99] and Theorem 6.6 of [Devroye and Lugosi 2001, p. 54] to obtain the L_1-error bound of the minimum distance estimate $f_{n-\varphi n,\theta^*}$, with $\theta^* \in \Theta$ and $|\Theta| < \infty$.

THEOREM 3.2. *If $\int f_{n-\varphi n,\theta} = 1$ for all $\theta \in \Theta$, then for the minimum distance estimate $f_{n-\varphi n,\theta^*}$ obtained by minimizing Δ_θ in* (3)*, we have*

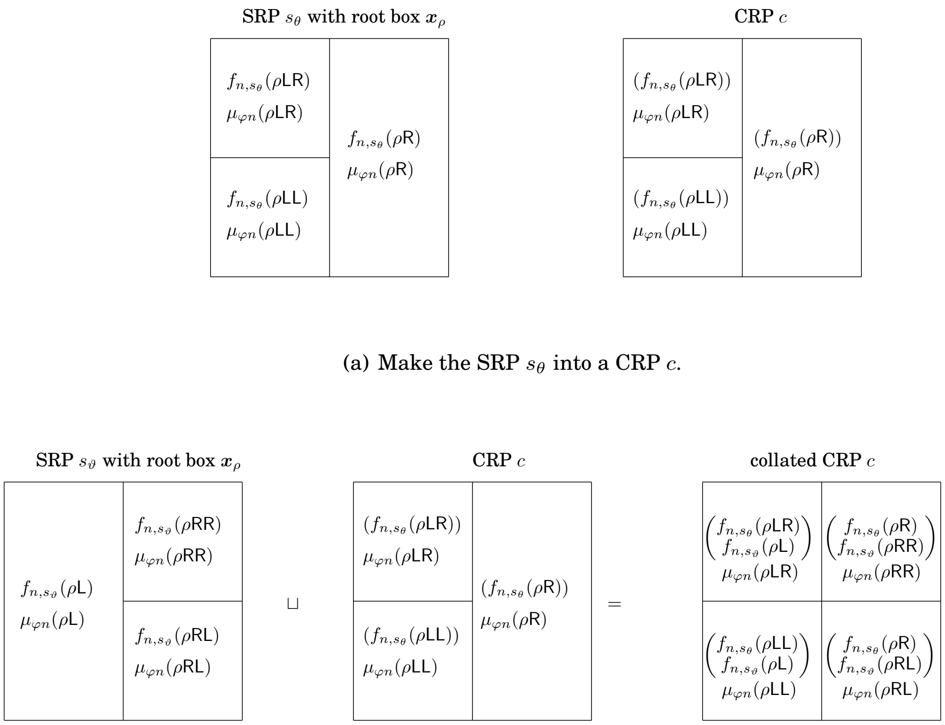
$$\int |f_{n-\varphi n,\theta^*} - f| \leq 3 \min_{\theta\in\Theta} \int |f_{n-\varphi n,\theta} - f| + 4\Delta \tag{4}$$

where

$$\Delta = \max_{A\in\mathcal{A}_\Theta} \left| \int_A f - \mu_{\varphi n}(A) \right| \ . \tag{5}$$

Theorem 3.2 can be proved directly by a conditional application of Theorem 6.3 of Devroye and Lugosi [2001, p. 54] and is nothing but the finite Θ version of their Theorem 10.1 [Devroye and Lugosi 2001, p. 99] without the additional $3/n$ term due to $|\Theta| < \infty$.

When f is unknown and $2^n > |\mathcal{A}_\Theta|$, Δ may be approximated by using the cardinality bound [Devroye et al. 1996, Theorem 13.6, p. 219] for the shatter coefficient of $\mathcal{A}_\Theta$.


SRP s_θ with root box $\boldsymbol{x}_\rho$
$f_{n,s_\theta}(\rho\mathsf{LR})$
$\mu_{\varphi n}(\rho\mathsf{LR})$
$f_{n,s_\theta}(\rho\mathsf{R})$
$\mu_{\varphi n}(\rho\mathsf{R})$
$f_{n,s_\theta}(\rho\mathsf{LL})$
$\mu_{\varphi n}(\rho\mathsf{LL})$
CRP c
$(f_{n,s_\theta}(\rho\mathsf{LR}))$
$\mu_{\varphi n}(\rho\mathsf{LR})$
$(f_{n,s_\theta}(\rho\mathsf{R}))$
$\mu_{\varphi n}(\rho\mathsf{R})$
$(f_{n,s_\theta}(\rho\mathsf{LL}))$
$\mu_{\varphi n}(\rho\mathsf{LL})$
SRP s_ϑ with root box $\boldsymbol{x}_\rho$
$f_{n,s_\vartheta}(\rho\mathsf{RR})$
$\mu_{\varphi n}(\rho\mathsf{RR})$
$f_{n,s_\vartheta}(\rho\mathsf{L})$
$\mu_{\varphi n}(\rho\mathsf{L})$
$f_{n,s_\vartheta}(\rho\mathsf{RL})$
$\mu_{\varphi n}(\rho\mathsf{RL})$
$\sqcup$
CRP c
$(f_{n,s_\theta}(\rho\mathsf{LR}))$
$\mu_{\varphi n}(\rho\mathsf{LR})$
$(f_{n,s_\theta}(\rho\mathsf{R}))$
$\mu_{\varphi n}(\rho\mathsf{R})$
$(f_{n,s_\theta}(\rho\mathsf{LL}))$
$\mu_{\varphi n}(\rho\mathsf{LL})$
$=$
collated CRP c
$\begin{pmatrix} f_{n,s_\theta}(\rho\mathsf{LR}) \\ f_{n,s_\vartheta}(\rho\mathsf{L}) \end{pmatrix}$
$\mu_{\varphi n}(\rho\mathsf{LR})$
$\begin{pmatrix} f_{n,s_\theta}(\rho\mathsf{R}) \\ f_{n,s_\vartheta}(\rho\mathsf{RR}) \end{pmatrix}$
$\mu_{\varphi n}(\rho\mathsf{RR})$
$\begin{pmatrix} f_{n,s_\theta}(\rho\mathsf{LL}) \\ f_{n,s_\vartheta}(\rho\mathsf{L}) \end{pmatrix}$
$\mu_{\varphi n}(\rho\mathsf{LL})$
$\begin{pmatrix} f_{n,s_\theta}(\rho\mathsf{R}) \\ f_{n,s_\vartheta}(\rho\mathsf{RL}) \end{pmatrix}$
$\mu_{\varphi n}(\rho\mathsf{RL})$

(a) Make the SRP s_θ into a CRP c.

(b) Collate another SRP s_ϑ onto CRP c.

Fig. 5. Collating two SRPs s_θ and s_ϑ with the same root box $\boldsymbol{x}_\rho$.

Given $\{x_1, \ldots, x_n\}$ the n-th shatter coefficient of $\mathcal{A}_\Theta$ is defined as

$$S\left(\mathcal{A}_\Theta, n\right) = \max_{x_1,\ldots,x_n \in \mathbb{R}^d} \left|\left\{\{x_1, \ldots, x_n\} \cap A : \quad A \in \mathcal{A}_\Theta\right\}\right| \quad .$$

Since $\mathcal{A}_\Theta$ is finite, containing at most quadratically many Scheffé sets $A_{\theta,\vartheta}$ with distinct ordered pairs $(\theta, \vartheta) \in \Theta^2$ given by the non-diagonal elements of the Yatracos matrix returned by `GetYatracos`, by Theorem 13.6 of Devroye et al. [1996, p. 219] its n-th shatter coefficient is bounded as follows:

$$S\left(\mathcal{A}_\Theta, n\right) \leq |\mathcal{A}_\Theta| \leq (|\Theta| + 1)^2 - (|\Theta| + 1) = |\Theta|(|\Theta| + 1) \quad . \tag{6}$$

Finally, given that adaptive multivariate histograms based on statistical regular pavings in $\mathbb{S}_{0:\infty}$ form a class of regular additive density estimates, we can slightly modify Theorem 10.3 of Devroye and Lugosi [2001, p. 103] for the case with finite Θ to get the following error bound that further accounts for splitting the data.

THEOREM 3.3. *Let $0 < \varphi < 1/2$ and $n < \infty$. Let the finite set Θ determine a class of adaptive multivariate histograms based on statistical regular pavings with $\int f_{n-\varphi n,\theta} = 1$ for all $\theta \in \Theta$. Let f_{n,θ^*} be the minimum distance estimate. Then for all n, φn, Θ and $f \in L_1$:*

$$E\left\{\int |f_{n-\varphi n,\theta^*} - f|\right\} \le 3\min_\theta E\left\{\int |f_{n,\theta} - f|\right\}\left(1 + \frac{2\varphi}{1-\varphi} + 8\sqrt{\varphi}\right) + 8\sqrt{\frac{\log 2|\Theta|(|\Theta|+1)}{\varphi n}} \ .$$

PROOF. By Theorem 3.2,

$$\int |f_{n-\varphi n,\theta^*} - f| \ \le \ 3\min_\theta \int |f_{n-\varphi n,\theta} - f| + 4\Delta$$

Taking expectations on both sides and using Theorem 10.2 in Devroye and Lugosi [2001, p. 99],

$$\begin{aligned} E\left\{\int |f_{n-\varphi n,\theta^*} - f|\right\} &\le \ 3\min_\theta E\left\{\int |f_{n-\varphi n,\theta} - f|\right\} + 4E\Delta \\ &\le \ 3\min_\theta E\left\{\int |f_{n,\theta} - f|\right\}\left(1 + \frac{2\varphi n}{(1-\varphi)n} + 8\sqrt{\frac{\varphi n}{n}}\right) + 4E\Delta \ . \end{aligned}$$

Finally by Theorem 3.1 in [Devroye and Lugosi 2001, p. 18] and (6),

$$\begin{aligned} 4E\Delta = 4E\left\{\sup_{A\in\mathcal{A}_\Theta}\left|\int_A f - \mu_{\varphi n}(A)\right|\right\} &\le \ 4\cdot 2\cdot\sqrt{\frac{\log 2S(\mathcal{A}_\Theta,\varphi n)}{\varphi n}} \\ &\le \ 4\cdot 2\cdot\sqrt{\frac{\log 2|\Theta|(|\Theta|+1)}{\varphi n}} \ . \end{aligned}$$

□

In order to effectively use the error bound we need to ensure that $|\Theta|$ is not too large and the densities in Θ are close to the true density f. Next, we highlight the effectiveness and limitations of our MDE.

The size of Θ is kept small (typically less than 100) and independent of n by an adaptive search. Note that $|\Theta|$ is upper-bounded by $\overline{m}$ if we were to exhaustively consider each SRP state along the entire path of the `SEBTreeMC` in Θ, our set of candidate SRP partitions. Such an exhaustive approach is computationally inefficient as the Yatracos matrix that updates the Scheffé sets grows quadratically with $|\Theta|$. We take a simple adaptive search approach by considering only k (typically $10 \le k \le 20$) SRP states in each iteration. In the initial iteration we add k states to Θ by picking uniformly spaced states from a long-enough `SEBTreeMC` path that starts from the root node and ends at a state with a large number of leaves and a significantly higher Δ_θ score than its preceding states. Then we simply zoom-in around the states with the lowest Δ_θ values and add another k states along the same `SEBTreeMC` path close to such optimal states from the first iteration. We repeat this adaptive search process until we are unable to zoom-in further. Typically, we are able to find nearly optimal states within 5 or fewer iterations. By Theorem 3.1, we know that the histogram partitioning strategy of `SEBTreeMC` is asymptotically consistent. Thus, the adaptive search set Θ that is selected iteratively from the set of histogram states along the path of `SEBTreeMC` with optimal Δ_θ values will naturally contain densities that approach f as n increases. However, the rate at which the L_1 distance between the best density in Θ and f approach 0 will depend on the complexity of f in terms of the number of leaves needed to uniformly approximate f using simple functions with SRP partitions, a class that is dense in

$\mathcal{C}(\boldsymbol{x}_\rho, \mathbb{R})$, the algebra of real-valued continuous functions over the root box $\boldsymbol{x}_\rho$ by the Stone-Weierstrass Theorem [Harlow et al. 2012, Theorem 4.1]. This dependence on the structural complexity of f is evaluated next.

4. PERFORMANCE EVALUATION

To evaluate the performance of our MDE we chose two multivariate densities: the spherically symmetric Gaussian and the highly structured Rosenbrock density (whose expression up to normalization is given in (7)) in d dimensions for various sample sizes.

$$\exp\left(-\sum_{i=2}^{d}(100(x_i - x_{i-1}^2)^2 + (1 - x_{i-1})^2)\right) \quad . \tag{7}$$

Table I. The MIAE for MDE and posterior mean estimates with different sample sizes for the 1D-, 2D-, and 5D-Gaussian densities, as well as the 2D- and 5D-Rosenbrock densities.

n	Standard Gaussian Densities			Rosenbrock Densities	
	1D	2D	5D	2D	5D
	Minimum Distance Estimate's Mean $L_1(f_{n,\theta^*}, f)$, $L_1(f_{n,\theta^*}, f) - \min_{\theta\in\Theta} L_1(f_{n,\theta}, f)$				
10^4	0.0888, 0.0058	0.2038, 0.0044	0.6764, 0.0020	0.4502, 0.0050	1.0154, 0.0018
10^5	0.0504, 0.0046	0.1140, 0.0014	0.4744, 0.0006	0.2476, 0.0024	0.7278, 0.0060
10^6	0.0204, 0.0014	0.0656, 0.0014	0.3310, 0.0006	0.1430, 0.0006	0.4772, 0.0034
10^7	0.0100, 0.0004	0.0376, 0.0002	0.2548, 0.0014	0.0828, 0.0012	0.2661, 0.0016
	MCMC Posterior Mean Estimate's MIAE (standard error)				
10^4	0.0565 (0.0053)	0.1673 (0.0046)	0.6467 (0.0051)	0.3717 (0.0103)	1.0190 (0.0059)
10^5	0.0274 (0.0011)	0.0932 (0.0002)	0.4655 (0.0020)	0.1982 (0.0067)	0.7250 (0.0011)
10^6	0.0129 (0.0006)	0.0533 (0.0005)	0.3274 (0.0009)	0.1102 (0.0006)	0.4812 (0.0012)
10^7	0.0060 (0.0001)	0.0304 (0.0002)	0.2292 (0.0034)	0.0608 (0.0049)	0.3302 (0.0004)

The sample standard deviations about the mean integrated absolute errors or MIAEs for the MDE method, i.e., $L_1(f_{n,\theta^*}, f)$ (shown in the top panel of Table I), based on ten trials, are below 10^{-3} and 10^{-4} for values of n in $\{10^4, 10^5\}$ and $\{10^6, 10^7\}$, respectively. Thus these standard errors are not shown. However, the L_1 distance between the MDE and the best estimate in the candidate set Θ, $L_1(f_{n,\theta^*}, f) - \min_{\theta\in\Theta} L_1(f_{n,\theta}, f)$, is shown in Table I for each density and sample size. For comparison we used the posterior mean histograms based on the MCMC method [Sainudiin et al. 2013, see for details on this evaluation] (they are shown in the bottom panel of Table I along with their standard errors. Note how the L_1 errors decrease with the sample size and how the errors are comparable between the methods, albeit the MDE method is at least an order of magnitude faster than the MCMC method (for detailed CPU times of the MCMC method see [Sainudiin et al. 2013]).

Remark 4.1. The approximate integration methods based on quasi-random streams and their importance sampling extensions became unreliable and significantly slower for highly structured densities such as that of Rosenbrock (7) in dimensions as large as 5. Thus, we used *r*eal mapped regular paving or $\mathbb{R}$-MRP approximation of the true density that is within 0.01 in Hellinger distance of the true density (see [Sainudiin et al. 2013, Sec. 4.2] and [Harlow et al. 2012] for details) whose n samples were simulated exactly using interval enclosures of the range of the target density [Sainudiin and York 2013] over regularly paved partitions. The target density f can be any one with a locally Lipschitz arithmetical expression and not merely the two examples shown here (see `mrs2` [Sainudiin et al. 2018] `examples/MooreRejSam` module) and this allows a skeptic to experiment for further evidence from simulations from this large class of densities on their own. By producing n samples from such piecewise constant $\mathbb{R}$-MRP densities, we can take advantage of $\mathbb{R}$-MRP arithmetic to obtain the exact L_1 error in Table I between the approximated $\mathbb{R}$-MRP representation of the density f and the $\mathbb{R}$-MRP representation of the estimate f_n produced by the MCMC or MDE methods. All experiments were performed on the same physical machine that is currently considered to be commodity hardware [Sainudiin et al. 2013, for machine specifications].

Thus, by using the collator regular paving (CRP), we obtain the minimum distance estimate (MDE) with universal performance guarantees. All the methods are implemented and available in `mrs2` [Sainudiin et al. 2018]. We limited our minimum distance estimate (MDE) to the candidate set given by the SRP histograms visited along the path of the Markov chain `SEBTreeMC`. This was done to take advantage of the the structure of consecutive refinements of the tree partitions along a single path of `SEBTreeMC`. However, obtaining the MDE from an arbitrary set of SRP histograms taken from $\mathbb{S}_{0:\infty}$ will need more sophisticated collators. Initial experiments using the Scheffé tournament approach (as opposed to the MDE) to find the best estimate in a candidate set of arbitrary SRP histograms (not just those along a path in $\mathbb{S}_{0:\infty}$) look feasible. Such a Scheffé tournament will allow us to compare estimates from entirely different methodological schools (Bayesian, penalized likelihood, etc.). Finally, the pure tree structure allows one to possibly extend this MDE to a distributed fault-tolerant computational setting such as Apache Spark [Zaharia et al. 2016] as the sample size becomes too large for the memory of a single machine.

ACKNOWLEDGMENTS

RS and GT proved the theorems and GT implemented the three MDE algorithms based on codes by Jennifer Harlow and RS in `mrs2`. This research began from a conversation RS had with Luc Devroye at the World Congress in Probability and Statistics in 2008 and was partly supported by RS's external consulting revenues from the New Zealand Ministry of Tourism, UC College of Engineering Sabbatical Grant and Visiting Scholarship at Department of Mathematics, Cornell University, Ithaca NY, USA and completed through the the project CORCON: Correctness by Construction, Seventh Framework Programme of the European Union, Marie Curie Actions-People, International Research Staff Exchange Scheme (IRSES) with counter-part funding from the Royal Society of New Zealand.

REFERENCES

DEVROYE, L., GYÖRFI, L., AND LUGOSI, G. 1996. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

DEVROYE, L. AND LUGOSI, G. 2001. *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York.

DEVROYE, L. AND LUGOSI, G. 2004. Bin Width Selection in Multivariate Histograms by the Combinatorial Method. *TEST 13,* 1, 129–145.

FISHER, R. A. 1925. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society 22*, 700–725.

GRAY, A. G. AND MOORE, A. W. 2003. Nonparametric Density Estimation: Towards Computational Tractability. In *SIAM International Conference on Data Mining*. SIAM, San Francisco, California, USA, 203–211.

HARLOW, J., SAINUDIIN, R., AND TUCKER, W. 2012. Mapped regular pavings. *Reliable Computing 16*, 252–282.

KIEFFER, M., JAULIN, L., BRAEMS, I., AND WALTER, E. 2001. Guaranteed set computation with subpavings. In *Scientific Computing, Validated Numerics, Interval Methods, Proceedings of SCAN 2000*, W. Kraemer and J. Gudenberg, Eds. Kluwer Academic Publishers, New York, 167–178.

KLEMELÄ, J. 2009. *Smoothing of Multivariate Data: Density Estimation and Visualization*. Wiley, Chichester, United Kingdom.

LU, L., JIANG, H., AND WONG, W. H. 2013. Multivariate density estimation by bayesian sequential partitioning. *Journal of the American Statistical Association 108,* 504, 1402–1410.

LUGOSI, G. AND NOBEL, A. 1996. Consistency of Data-Driven Histogram Methods for Density Estimation and Classification. *The Annals of Statistics 24,* 2, 687–706.

MAHALANABIS, S. AND STEFANKOVIC, D. 2008. Density estimation in linear time. In *21st Annual Conference on Learning Theory - COLT 2008*, R. A. Servedio and T. Zhang, Eds. Omnipress, Helsinki, Finland, 503–512.

MATTAREI, S. 2010. Asymptotics of partial sums of central binomial coefficients and Catalan numbers. arXiv.0906.4290v3.

MEIER, J. 2008. *Groups, Graphs and Trees: An Introduction to the Geometry of Infinite Groups*. Cambridge University Press, Cambridge, United Kingdom.

SAINUDIIN, R., TENG, G., HARLOW, J., AND LEE, D. S. 2013. Posterior expectation of regularly paved random histograms. *ACM Transactions on Modeling and Computer Simulation 23,* 26, 6:1–6:20.

SAINUDIIN, R. AND YORK, T. 2013. An auto-validating, trans-dimensional, universal rejection sampler for locally Lipschitz arithmetical expressions. *Reliable Computing 18*, 15–54.

SAINUDIIN, R., YORK, T., HARLOW, J., TENG, G., TUCKER, W., AND GEORGE, D. 2008–2018. MRS 2.0, a C++ class library for statistical set processing and computer-aided proofs in statistics. `https://github.com/raazesh-sainudiin/mrs2`.

SAMET, H. 1990. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley Longman, Boston.

STANLEY, R. P. 1999. *Enumerative combinatorics. Vol. 2*. Cambridge Studies in Advanced Mathematics Series, vol. 62. Cambridge University Press, Cambridge.

TUKEY, J. W. 1947. Non-Parametric Estimation II. Statistically Equivalent Blocks and Tolerance Regions — The Continuous Case. *The Annals of Mathematical Statistics 18,* 4, 529–539.

VAPNIK, V. N. AND CHERVONENKIS, A. Y. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl. 16*, 264–280.

YATRACOS, Y. G. 1985. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics 13,* 2, pp. 768–774.

YATRACOS, Y. G. 1988. A note on l1 consistent estimation. *The Canadian Journal of Statistics 16,* 3, 283–292.

ZAHARIA, M., XIN, R. S., WENDELL, P., DAS, T., ARMBRUST, M., DAVE, A., MENG, X., ROSEN, J., VENKATARAMAN, S., FRANKLIN, M. J., GHODSI, A., GONZALEZ, J., SHENKER, S., AND STOICA, I. 2016. Apache Spark: A unified engine for big data processing. *Commun. ACM 59,* 11, 56–65.

Appendix: MDE Algorithms

ALGORITHM 2: `GetDelta`

```
input     :
(1) the current number of splits: $i$;
(2) the collated regular paving CRP: $c$ with pointers to the vector $\boldsymbol{f}_{n-\varphi n,c}(\rho)$ and
    $\mu_{\varphi n}(\rho)$ of each node in $c$
(3) the Yatracos matrix: $\mathcal{A}_{\Theta_i}$;
(4) the current $\Delta_\theta$ vector: $\Delta_{\Theta_{i-1}} \in \mathbb{R}^{(1\times(i))}$.

output    : the updated $\Delta_\theta$ vector: $\Delta_{\Theta_i} \in \mathbb{R}^{(1\times(i+1))}$.

if $i = 0$ then
    $\Delta_{\Theta_i} = \emptyset$
end
else
    // Get $\Delta_\theta$ for all $\theta \in \Theta_{i-1}$ for the sets in the $(i+1)$-column and the
        $(i+1)-th$ row of $\mathcal{A}_{\Theta_i}$.
    foreach $\theta \in \Theta_{i-1}$ do
        foreach $A \in \{\mathcal{A}_{\Theta_i}(\cdot, i+1), \mathcal{A}_{\Theta_i}(i+1, \cdot)\}$ do
            $\Delta \leftarrow 0$
            foreach $\boldsymbol{x} \in A$ do
                $\Delta \leftarrow \Delta + \left[\left(\boldsymbol{f}^{(\theta)}_{n-\varphi n,c}(\boldsymbol{x}) * \mathrm{vol}\,(\boldsymbol{x})\right) - \mu_{\varphi n}(\boldsymbol{x})\right]$
            end
            $\Delta \leftarrow |\Delta|$
            $\Delta_\theta \leftarrow \max\{\Delta, \Delta_\theta\}$
        end
        insert $\Delta_\theta$ into $\Delta_{\Theta_i}(\theta)$ ;   // insert into the $\theta$-th entry of the vector $\Delta_{\Theta_i}$
    end

    // Get $\Delta_\theta$ for $\theta = i$
    foreach $A \in \{\mathcal{A}_{\Theta_i}$ do
        $\Delta \leftarrow 0$
        foreach $\boldsymbol{x} \in A$ do
            $\Delta \leftarrow \Delta + \left[\left(\boldsymbol{f}^{(\theta)}_{n-\varphi n,c}(\boldsymbol{x}) * \mathrm{vol}\,(\boldsymbol{x})\right) - \mu_{\varphi n}(\boldsymbol{x})\right]$
        end
        $\Delta \leftarrow |\Delta|$
        $\Delta_\theta \leftarrow \max\{\Delta, \Delta_\theta\}$
    end
    insert $\Delta_\theta$ into $\Delta_{\Theta_i}(i+1)$
end
return $\Delta_{\Theta_i}$
```

ALGORITHM 3: `SRPCollate`$(\rho, \rho^{(c)})$

```
input      :
(1) The root node $\rho$ of an SRP $s$ with root box $\boldsymbol{x}_\rho$.
(2) The root node $\rho^{(c)}$ of an CRP $c$.
output     : The updated root node $\rho^{(c)}$ of the CRP $c$.
if $\rho^{(c)} = \emptyset$ // Nothing has been collated yet.
 then
    Make a new node $\rho^{(c)}$ with box $\boldsymbol{x}_\rho$
    foreach $\rho v \in s$ do
        $f_{n-\varphi n,s}(\rho v) \leftarrow \#\boldsymbol{x}_{\rho v}/((n-\varphi n) * \rho v)$
        Insert $f_{n-\varphi n,s}(\rho v)$ into $\boldsymbol{f}_{n-\varphi n,c}(\rho v)$ ;    // This is a ‘‘pushback’’ operation,
         i.e keep $f_{n-\varphi n,s}(\rho v)$ in a vector $\boldsymbol{f}_{n-\varphi n,c}(\rho v)$.
        $\mu_{\varphi n}(\rho v) \leftarrow \check{\#}\boldsymbol{x}_{\rho v}/\varphi n$
    end
    return $c$
end
else
    Make a new node $\rho^{(c)}$ with box $\boldsymbol{x}_\rho$
    $f_{n-\varphi n,s}(\rho^{(c)}) \leftarrow \#\boldsymbol{x}_{\rho^{(c)}}/(n * \rho^{(c)})$
    Insert $f_{n-\varphi n,s}(\rho^{(c)})$ into $\boldsymbol{f}_{n-\varphi n,c}(\rho)$
    $\mu_{\varphi n}(\rho^{(c)}) \leftarrow \check{\#}\boldsymbol{x}_{\rho^{(c)}}/\varphi n$

    if (IsLeaf($\rho$) & (!IsLeaf($\rho^{(c)}$)) then
        Make temporary nodes L′, R′
        $\boldsymbol{x}_{\mathsf{L}'} \leftarrow \boldsymbol{x}_{\rho\mathsf{L}}$, $\boldsymbol{x}_{\mathsf{R}'} \leftarrow \boldsymbol{x}_{\rho\mathsf{R}}$
        $f_{n-\varphi n,s}(\mathsf{L}') \leftarrow f_{n-\varphi n,s}(\rho)$, $f_{n-\varphi n,s}(\mathsf{R}') \leftarrow f_{n-\varphi n,s}(\rho)$
        Graft onto $\rho^{(c)}$ as left child the node SRPCollate($\mathsf{L}', \rho^{(c)}\mathsf{L}$)
        Graft onto $\rho^{(c)}$ as right child the node SRPCollate($\mathsf{R}', \rho^{(c)}\mathsf{R}$)
    end

    if (IsLeaf($\rho^{(c)}$) & (!IsLeaf($\rho$) then
        Make temporary nodes L′, R′
        $\boldsymbol{x}_{\rho\mathsf{L}'} \leftarrow \boldsymbol{x}_{\rho^{(c)}\mathsf{L}}$, $\boldsymbol{x}_{\mathsf{R}'} \leftarrow \boldsymbol{x}_{\rho^{(c)}\mathsf{R}}$
        $f_{n-\varphi n,s}(\mathsf{L}') \leftarrow f_{n-\varphi n,s}(\rho^{(c)})$, $f_{n-\varphi n,s}(\mathsf{R}') \leftarrow f_{n-\varphi n,s}(\rho^{(c)})$
        Graft onto $\rho^{(c)}$ as left child the node SRPCollate($\rho\mathsf{L}, \mathsf{L}'$)
        Graft onto $\rho^{(c)}$ as right child the node SRPCollate($\rho\mathsf{R}, \mathsf{R}'$)
    end

    if (!IsLeaf($\rho$)) & (!IsLeaf($\rho^{(c)}$) then
        Graft onto $\rho^{(c)}$ as left child the node SRPCollate($\rho\mathsf{L}, \rho^{(c)}\mathsf{L}$)
        Graft onto $\rho^{(c)}$ as right child the node SRPCollate($\rho\mathsf{R}, \rho^{(c)}\mathsf{R}$)
    end
    return $\rho^{(c)}$
end
```

ALGORITHM 4: `GetYatracos`

```
input    :
(1) the node that was split: $\rho v^*$;
(2) the vector of histogram estimates: $\boldsymbol{f}_{n-\varphi n,c}$;
(3) the current number of splits: $i$;
(4) the current Yatracos matrix: $\mathcal{A}_{\Theta_{i-1}}$.
output   : the updated Yatracos matrix: $\mathcal{A}_{\Theta_i}$.
if $\boldsymbol{x}_{\rho v^*} = \boldsymbol{x}_\rho$ then
    $A_{0,0} \leftarrow \emptyset$
end
for $j = 0 : (i-1)$ do
    check the i-th column // Iterating through the entries of the $(i-1)$-th
        column to check if the entry $A_{j,i-1}$ contains $\boldsymbol{x}_{\rho v^*}$
    if $(A_{j,i-1} \neq \emptyset)$ & $(\boldsymbol{x}_{\rho v^*} \in A_{j,i-1})$ then
        // The entry $A_{j,i}$ takes all the elements of $A_{j,i-1}$ except $\boldsymbol{x}_{\rho v^*}$
        $A_{j,i} \leftarrow A_{j,i-1} \setminus \boldsymbol{x}_{\rho v^*}$
    end
    else
        $A_{j,i} \leftarrow A_{j,i-1}$
    end
    // Compare the estimates at each child node
    foreach $\boldsymbol{x} \in \{\boldsymbol{x}_{\rho v^*\mathsf{L}}, \boldsymbol{x}_{\rho v^*\mathsf{R}}\}$ do
        if $\boldsymbol{f}^{(j)}_{n-\varphi n,c}(\boldsymbol{x}_\rho) > \boldsymbol{f}^{(i)}_{n-\varphi n,c}(\boldsymbol{x}_\rho)$ then
            // Take the union of the elements in entry $A_{j,i}$ with $\boldsymbol{x}_\rho$
            $A_{j,i} \leftarrow \left\{ \bigcup_{\boldsymbol{x}_v \in A_{j,i}} \boldsymbol{x}_{\rho v} \cup \boldsymbol{x}_\rho \right\}$
        end
    end
    check the i-th row // Iterating through the entries of the $(i-1)$-th row to
        check if the entry $A_{i-1,j}$ contains $\boldsymbol{x}_{\rho v^*}$
    if $(A_{i-1,j} \neq \emptyset)$ & $(\boldsymbol{x}_{\rho v^*} \in A_{i-1,j})$ then
        // The entry $A_{i,j}$ takes all the elements of $A_{i-1,j}$ except $\boldsymbol{x}_{\rho v^*}$
        $A_{i,j} \leftarrow A_{i-1,j} \setminus \boldsymbol{x}_{\rho v^*}$
    end
    else
        $A_{i,j} \leftarrow A_{i-1,j}$
    end
    // Compare the estimates at each child node
    foreach $\boldsymbol{x}_\rho \in \{\boldsymbol{x}_{\rho v^*\mathsf{L}}, \boldsymbol{x}_{v^*\mathsf{R}}\}$ do
        if $\boldsymbol{f}^{(i)}_{n-\varphi n,i}(\boldsymbol{x}_\rho) > \boldsymbol{f}^{(j)}_{n-\varphi n,j}(\boldsymbol{x}_\rho)$ then
            // Take the union of the elements in entry $A_{i,j}$ with $\boldsymbol{x}_\rho$
            $A_{i,j} \leftarrow \left\{ \bigcup_{\boldsymbol{x}_{\rho v} \in A_{i,j}} \boldsymbol{x}_{\rho v} \cup \boldsymbol{x}_\rho \right\}$
        end
    end
end
$A_{i,i} \leftarrow \emptyset$ // The diagonal entry is always an empty set
return $\mathcal{A}_{\Theta_i}$
```