Density estimation using cellular binary trees and an application to monotone densities

Luc Devroye and Jad Hamdan

School of Computer Science, McGill University

and

Department of Mathematics, Oxford University

Abstract. Consider a density f on $[0,1]$ that must be estimated from an i.i.d. sample $X_1, ..., X_n$ drawn from f. In this note, we study binary-tree-based histogram estimates that use recursive splitting of intervals. If the decision to split an interval is a (possibly randomized) function of the number of data points in the interval only, then we speak of an estimate of complexity one. We exhibit a universally consistent estimate of complexity one. If the decision to split is a function of the cardinalities of k equal-length sub-intervals, then we speak of an estimate of complexity k. We propose an estimate of complexity two that can estimate any bounded monotone density on $[0,1]$ with optimal expected total variation error $O(n^{-1/3})$.

Keywords. Density estimation, monotone densities, nonparametric estimation, cellular computation, binary trees, Galton–Watson trees.

MSC2020 subject classification. Primary 62G07, 68Q87, 68W40; secondary 60C05, 60J85.

1. Introduction

We are concerned with the estimation of an unknown density f on $[0,1]$ based on an i.i.d. sample $X_1, ..., X_n$ drawn from it. In particular, given a recursive partition of the space into intervals, one can simply use the partition-based histogram estimate: on a fixed interval C, f is estimated by

$$f_n(x) = \frac{N(C)}{n\lambda(C)}, \quad x \in C,$$

where C is the unique interval to which x belongs, $\lambda(C)$ is the length (or Lebesgue measure) of C, and $N(C)$ is the number of X_i's falling in C. We impose two further design restrictions, one for convenience, and one motivated by distributed computation.

The only partitions of $[0,1]$ allowed are dyadic. This lets us view each of the allowed partitions as a binary tree. The root represents $[0,1]$, the two children

of the root represent $[0, 1/2]$ and $[1/2, 1]$, and at level i in the tree, we have an equi-partition of $[0, 1]$ into 2^i intervals of length $1/2^i$. The leaves in the tree correspond to intervals whose union is $[0, 1]$. (We abuse the term partition, as we allow intervals to overlap in a border point.) To estimate f, it suffices to construct this tree given the data. The data mining community (Schmidberger, 2009 [14]; Ram and Gray, 2011 [13]; Anderlini, 2016 [2]) refers to this general method as "density estimation trees". We note that in the event where the tree is not full, the resulting partition's bins can have different lengths.

Our second restriction is motivated by distributed computation. At the root, we may decide to either split the root interval or make it a leaf. If it is split, the data travel to their respective sub-trees. So, the data flow down the tree according to where they belong. Every node in the tree must act similarly, so we will refer to nodes as cells. The decision to split has a complexity parameter κ. When $\kappa = 1$, the decision can only depend upon N, the number of points that fall in the node's interval, C. In particular, that decision can't depend upon the original n, and can be handled by an autonomous computer, or "cell". For $\kappa > 1$, the decision only depends upon $N_1, \ldots, N_k$, the cardinalities of the k equal-length sub-intervals of C. This view was also eschewed by Biau and Devroye (2013) [5].

Many things can go awry. For one thing, we could end up with an infinite tree without any leaves. Or we could end up with just one node, the root. In the former case, the estimate is not defined. In the latter case, one gets consistency only if f itself is uniform $[0, 1]$. The study here is a small first step into this new model of "cellular" estimators (meaning those which satisfy our restrictions), noting that generalizations in many directions are possible. Most importantly, one could consider 2^d-ary trees to partition $[0, 1]^d$ for general d. Computing will likely become more distributed and miniaturized, making our model relevant.

We begin with the description of an estimate of complexity one that is universally consistent, i.e., for all densities f on $[0, 1]$,

$$\int |f_n - f| \to 0$$

in probability as $n \to \infty$.

Then we exhibit an estimate of complexity two that consistently estimates every bounded monotone decreasing density on $[0, 1]$ at an optimal rate in n. The decision to split is made when $N_1 - N_2 > \gamma\sqrt{N_1 + N_2}$ for fixed universal design constant $\gamma > 0$. We will show that

$$\mathbf{E}\left\{ \int_0^1 |f_n(x) - f(x)|\, dx \right\} = O\left(\frac{B^{2/3}}{n^{1/3}}\right)$$

where $B = f(0)$ is the value at the mode, and f_n is the histogram estimate on the partition induced by the leaves. Also, when $B = \infty$, we still have consistency, i.e., $\mathbf{E}\{\int |f_n - f|\} \to 0$ as $n \to \infty$.

It is noteworthy that if $\mathcal{M}_B$ denotes the class of all monotone densities on $[0, 1]$ bounded by B, then

$$\inf_{\text{all estimators } f_n} \sup_{f \in \mathcal{M}_B} \mathbf{E}\left\{ \int_0^1 |f_n(x) - f(x)|\, dx \right\} \geq \alpha \left(\frac{\log(1+B)}{n} \right)^{1/3}$$

for a universal constant $\alpha > 0$ (Birgé, 1983 [6]; see also Devroye and Györfi, 1985 [8]). Our estimate achieves the minimax rate in n, albeit with a suboptimal constant multiplicative factor. Several estimates achieve the minimax rate with the correct multiplicative factor, notably Grenander's histogram estimate (Grenander, 1956 [9]) (which uses a partition based on the smallest concave majorant of the distribution function) and Birgé's histogram estimate (Birgé, 1983 [6]) (which uses a partition of exponentially increasing widths). Standard histogram estimates of equal bin widths can at best achieve a rate proportional to $(B/n)^{1/3}$.

Our simple estimate does not use additional information that one may know about the density. Prior knowledge of the smoothness, for instance, is completely ignored. With explicit knowledge of smoothness in terms of Hölder coefficient $\beta \geq 1$, the minimax rate can be of the order of $n^{-\beta/(2\beta+1)}$, surpassing $n^{-1/3}$ if $\beta > 1$. This rate can be achieved with the kernel density estimate (Wasserman, 2006 [15]).

This paper aims to present the most straightforward and most general estimate using the design restriction outlined above, to showcase what can still be achieved despite this hurdle. For this reason, we do not give much importance to the smoothness issue discussed in the previous paragraph. After stating the main results announced above, we discuss the size of the tree obtained for the monotone density estimate and address the computational complexity. We note the importance of Galton–Watson trees in the analysis: as every density is locally nearly uniform, the performance of our splitting rule on a uniform density f explains the behaviour near the bottom of the tree. We will show for example that for uniform f, the binary tree is essentially an extinct Galton–Watson tree of constant expected size. We end the paper with extensions of the design principle to estimate densities with special structures such as convex or concave densities, log-concave or log-convex densities, and unimodal densities spring to mind.

2. A universally consistent estimate of complexity one

Any deterministic splitting rule of complexity $\kappa = 1$ is doomed because one can't decide which number of points $N(C)$ in a given interval is large enough to stop splitting. However, randomization can be used in the design. Assume that we have a non-increasing function $\varphi : Z \to [0, 1]$, and that our estimate of complexity one is:

$$\text{do not split } C \text{ when } U \leq \varphi(N),$$

where $N = N(C)$ is the cardinality of the interval C and U is an independent uniform $[0, 1]$ random variable. In this case, we obtain universal consistency:

Theorem 1. *Let f be a probability density function on $[0, 1]$. Then*

$$\int_0^1 |f_n - f| \to 0$$

in probability as $n \to \infty$, provided that

$$\lim_{n\to\infty} \varphi(n) = 0$$

and

$$\lim_{n\to\infty} \varphi(n) \log_2(n) = \infty.$$

Proved in the Appendix, Theorem 1 raises new questions regarding the tree's size, which measures the total computation time. One would also need information on the expected number of steps required to find the partition to which a point x belongs, as that would be proportional to the expected time to compute the density estimate at one point. In addition, the height of the tree would be of interest. Finally, the choice of φ within the bounds outlined in Theorem 1 should be studied.

3. An estimate of complexity two for monotone densities

When f is monotone on $[0, 1]$ and non-increasing, an interval C with equal-length sub-intervals C_1 and C_2 of cardinalities N_1 and N_2 can be split by the following rule of complexity two:

$$\text{split } C \text{ when } N_1 - N_2 > \gamma\sqrt{N_1 + N_2}$$

for fixed universal design constant $\gamma > 1$, noting that one would expect $N_1 \geq N_2$ by monotonicity of f. For the resulting estimate f_n we obtain an explicit upper bound on the total variation error:

Theorem 2. *Let f be a bounded non-increasing probability density function on $[0,1]$ with $B = f(0)$. Then*

$$\sup_{f \in \mathcal{M}_B} \mathbf{E}\left\{ \int_0^1 |f_n(x) - f(x)| dx \right\} \leq \beta \left(\frac{B^{2/3}}{n^{1/3}} \right)$$

for a universal constant β and large enough n.

Remark 1. Note that if we partition $[0,1]$ into k equal intervals, and let f_n be the standard histogram estimate for these k intervals, then

$$\mathbf{E}\left\{ \int_0^1 |f_n - f| \right\} \leq \mathbf{E}\left\{ \int_0^1 |f_n - \mathbf{E}f_n| \right\} + \int_0^1 |\mathbf{E}f_n - f|,$$

where $\mathbf{E}\{f_n(x)\} = p(C)/\lambda(C)$, $x \in C$, and $p(C) = \int_C f$. Thus, $\mathbf{E}\{f_n(x)\} = kp(C)$. A simple shifting argument shows that

$$\int_0^1 |\mathbf{E}f_n - f| \leq \frac{B}{k}.$$

Also,

$$\begin{aligned}
\mathbf{E}\left\{ \int_0^1 |f_n - \mathbf{E}f_n| \right\} &= \sum_C \mathbf{E}\left\{ \left| \frac{N(C)}{n} - p(C) \right| \right\} \\
&\leq \sqrt{\sum_C 1 \cdot \sum_C \mathbf{E}\left\{ \left| \frac{N(C)}{n} - p(C) \right|^2 \right\}} \\
&\leq \sqrt{\frac{k}{n} \sum_C p(C)} = \sqrt{\frac{k}{n}},
\end{aligned}$$

and therefore, taking $k = \lceil (2B)^{2/3} n^{1/3} \rceil$ to optimize the sum, we obtain

$$\sup_{f \in \mathcal{M}_B} \mathbf{E}\left\{ \int_0^1 |f_n - f| \right\} \leq \left(\frac{1}{2^{2/3}} + 2^{1/3} + o(1) \right) \left(\frac{B}{n} \right)^{1/3}.$$

Note that without knowledge of B, the histogram estimate does not have a better convergence rate than our handicapped estimate.

4. Monotone density estimate: algorithm and time complexity

From an algorithmic standpoint, the splitting described above amounts to a branching process that constructs a binary tree. For any $x \in \mathbf{R}$ and sorted list of numbers L, let $i(L, x)$ denote the index of x if it were inserted into L. Given a sorted list of size n whose elements are an i.i.d. sample $X_1, ..., X_n$ drawn from an unknown density f on $[a, b] \subseteq \mathbf{R}$, the following recursive algorithm constructs a partition tree of $[a, b]$ according to our splitting rule.

While not strictly necessary, the assumption that $[X_1, ..., X_n]$ is sorted allows us to decide whether or not to split in logarithmic time by using binary search to

Algorithm 1 Interval partitioning using a binary tree

function BUILDTREE(r, $[X_1, ..., X_n], [a, b]$) ▷ r is a tree node
  $L \leftarrow i([X_1, ..., X_n], (a+b)/2)$ ▷ L is the number of data points in the left half of $[a, b]$
  $R \leftarrow n - L$ ▷ R is the number of data points on the right half of $[a, b]$
  if $L - R > \gamma\sqrt{n}$ **then** ▷ γ is a parameter in $(0, \infty)$
    BUILDTREE(r.left, $[X_1, ..., X_L], [a, (a+b)/2]$)
    BUILDTREE(r.right, $[X_{L+1}, ..., X_n], [(a+b)/2, b]$)
  else
    r.value $\leftarrow [a, b]$
  end if
end function
Initialize new tree node r
BUILDTREE($r, [X_1, ..., X_n], [a, b]$)
return r

compute the number of points on the left and right halves of $[a, b]$. It also greatly simplifies the algorithm's pseudo-code to construct left and right sub-lists when we perform a recursive call.

As a corollary to the results shown later in the paper, we can derive the following sub-linear upper bound on the expected runtime of our algorithm.

Corollary 3. *This algorithm's expected runtime is* $O(n^{1/3}\log_2(n))$ *if the input data are sorted.*

Proof. See appendix. ■

5. Monotone density estimate: Galton–Watson trees and the uniform case

We recall the definition of a Galton–Watson tree (see, e.g., Athreya and Ney, 1972 [3]): the number of offspring of each node in the tree is random and distributed as Z, where $Z \geq 0$ has a fixed distribution. All realizations of Z are independent. If $\mathbf{E}Z = m < 1$, then the expected size of the tree is $1/(1-m)$. See, e.g., Lyons and Peres, 2016 [11].

As previously discussed, our splitting procedure can be viewed as a (randomly generated) binary tree of intervals which we will henceforth denote by T_n. An elegant connection to the theory of branching processes can be established when our data are sampled from a uniform distribution. More specifically, one can show that the resulting tree would closely resemble a Galton–Watson tree

whose nodes have two children with probability $p_2 = \mathbf{P}\{\mathcal{N}(0,1) > \gamma\} := \Phi(\gamma)$ and no children with probability $p_0 = 1 - p_2$ (where $\mathcal{N}(0,1)$ is a standard normal and γ is the parameter chosen in the algorithm, which is assumed to be ≥ 1 in this section).

Let C be an arbitrary sub-interval of $[0,1]$ and assume that C contains N data points. The number of points in the left and right halves of C, denoted by N_1 and N_2, respectively, are binomial random variables with parameters N and $1/2$. Noting that $2N_1 - N = N_1 - N_2$, the probability of splitting the interval is

$$\mathbf{P}\{N_1 - N_2 > \gamma\sqrt{N}\} = \mathbf{P}\left\{\frac{N_1 - N/2}{\sqrt{N/4}} > \gamma\right\},$$

which by the Berry-Esseen theorem (Berry, 1941 [4], see also Petrov, 1975 [12]) is equal to $\Phi(\gamma) + \theta/\sqrt{N} =: p_2$ for some $|\theta| \leq 1$.

Let $\epsilon > 0$ be arbitrary, and let T_n' be the subtree of T_n in which all nodes C (we refer to C as a node as well as an interval associated with that node) contain at least $N_\epsilon := \lceil 1/(\epsilon \cdot \Phi(\gamma))^2 \rceil$ points. Then for these nodes, the probability p_2 of splitting is smaller than

$$(1 + \theta \cdot \epsilon)\Phi(\gamma).$$

We infer that for ϵ small enough,

$$\mathbf{E}\{|T_n'|\} \leq \frac{1}{1 - 2(1+\epsilon)\Phi(\gamma)}.$$

Furthermore, every leaf of T_n' is either a leaf of T_n or an internal node containing less than N_ϵ points. In the latter case, we can derive a uniform upper bound for the expected size of sub-trees that hang from such leaves as a function of ϵ.

Assuming $\gamma \geq 1$, our splitting criterion is such that any interval with a single point is never split. By analyzing the expected minimum distance between any two points in an interval, we can determine an upper bound for the expected height (and in turn size) of a tree.

Consider N_ϵ uniformly distributed points on an interval (without loss of generality, $[0,1]$). Let D be an integer random variable taking value i when the minimum distance between two points of the interval lies in $(2^{-i-1}, 2^{-i}]$.

We split $[0,1]$ dyadically until each interval has 0 or 1 point (as depicted in figure 1). The expected number of internal nodes of this tree is

$$\sum_{\ell=0}^{\infty} 2^\ell \cdot \mathbf{P}\left\{\left[0, \frac{1}{2^\ell}\right] \text{ contains at least 2 points}\right\}$$
$$\leq \sum_{\ell=0}^{\infty} 2^\ell \cdot \binom{N_\epsilon - 1}{2} \frac{1}{2^{2\ell}} \leq \frac{(N_\epsilon - 1)^2}{2} \sum_{\ell=0}^{\infty} \frac{1}{2^\ell} = (N_\epsilon - 1)^2$$

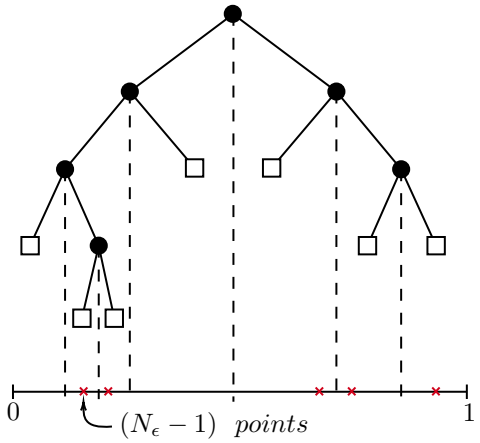

Figure 1. Dyadic splitting until each interval contains at most one point.

where ℓ is the level number in the tree.

Thus, the expected size is $\leq 2(N_\epsilon - 1)^2 + 1 < 2N_\epsilon^2$ since the number of leaves equals the number of internal nodes plus one.

We conclude that, under the assumption of uniformly distributed data, the expected tree size is finite and uniformly bounded over all values of n:

$$\begin{aligned}\mathbf{E}\{|T_n|\} &\leq \inf_{\epsilon>0} \frac{1}{1-2(1+\epsilon)\Phi(\gamma)} \cdot 2N_\epsilon^2 \\ &\leq \inf_{\epsilon>0} \frac{2}{1-2(1+\epsilon)\Phi(\gamma)} \cdot \left(\frac{1}{\left(\epsilon \cdot \Phi(\gamma)\right)^2} + 1 \right)^2 \stackrel{\text{def}}{=} \varphi(\gamma) < \infty.\end{aligned}$$

Similar reasoning yields the following theorem.

Theorem 4. *Let $f = 1$ on $[0,1]$ and $f = 0$ elsewhere. Then if $\Phi(\gamma) + \frac{1}{\gamma} < 1/2$,*

$$\mathbf{E}\left\{ \int_0^1 |f_n(x) - f(x)| dx \right\} = O\left(\frac{1}{\sqrt{n}} \right).$$

Remark 2. Any choice of $\gamma \geq 3$ ensures that this condition is satisfied.

Proof. Fix $x \in [0,1]$. Conditioning on the height h of the leaf to which x belongs (in the partition tree defining f_n) and using Cauchy-Schwarz, we get

$$\begin{aligned}\mathbf{E}\{|f_n(x) - f(x)|\} &\leq \sum_{\ell \geq 0} \sqrt{\mathbf{P}\{h = \ell\}\, \mathbf{E}\{|f_n(x) - f(x)|^2 \,|\, h = \ell\}} \\ &= \frac{1}{\sqrt{n}} \sum_{\ell \geq 0} \sqrt{2^\ell \mathbf{P}\{h = \ell\}}.\end{aligned}$$

We know that for any node containing N points, the probability of it splitting is bounded by $p_N = \Phi(\gamma) + 1/\sqrt{N}$. However, the condition on γ implies that

$p_N < 1/2$ uniformly. It follows that $\mathbf{P}\{h = \ell\} < (1/2)^\ell$, and the summation above is $O(1)$ as a geometric series. Applying Tonelli's theorem to $\mathbf{E}\{\int |f_n - f|\}$ thus yields the desired result. ■

6. The deterministic infinite tree

6.1. Notation, setup and main proposition

Towards our goal of proving Theorem 2, we begin with the analysis of the infinite full binary tree depicted in figure 3 and denoted by $\mathcal{T}_\infty$. It is analogous to T_n in that each node of $\mathcal{T}_\infty$ is associated with a sub-interval of $[0, 1]$; more specifically, if $C_1, ..., C_{2^\ell}$ are $\mathcal{T}_\infty$'s level ℓ nodes labelled left to right, then C_i corresponds to the interval $[(i-1)/2^\ell, i/2^\ell]$ for any $1 \le i \le 2^\ell$. It helps to view the random tree T_n as a subset of this infinite deterministic tree.

As in the introduction, the left and right halves of a node $C \in \mathcal{T}_\infty$ are denoted by C_1 and C_2, respectively. It is said to be *balanced* (and is uncoloured in figure 3) if it satisfies

$$p(C_1) - p(C_2) \le \gamma \sqrt{\frac{p(C)}{n}}, \tag{1}$$

where γ is the parameter previously defined for the algorithm (0) and, as above, $p(C) = \int_C f$. The set of all such nodes is denoted by $\mathcal{B}$. All other (coloured) nodes are said to be *unbalanced* and belong to $\mathcal{B}^c$, the complement of $\mathcal{B}$. Similarly, for any positive real number α, we denote by $\mathcal{B}^{(\alpha)}$ the set of nodes satisfying

$$p(C_1) - p(C_2) \le \alpha\gamma \sqrt{\frac{p(C)}{n}},$$

noting that $\mathcal{B} = \mathcal{B}^{(1)}$. The integer ℓ^* is defined as

$$\ell^* = \min \left\{ \ell \in \mathbf{Z}_{>0} : \frac{B}{2^{\ell+1}} \le \gamma \cdot 2^{\ell/2} \sqrt{\frac{B}{n}} \right\}.$$

We denote by $\mathcal{P}_j(\mathcal{T}_\infty) = \mathcal{P}_j$ the set of nodes of $\mathcal{T}_\infty$ with *exactly* j balanced ancestors. Lastly, for any node C, the average value of f on C is denoted by $f(C)$.

Note that if we were to truncate $\mathcal{T}_\infty$ by deleting nodes that fall below those belonging to $\mathcal{B} \cap \mathcal{P}_0$, the resulting tree (with leaf set $\mathcal{B} \cap \mathcal{P}_0$) would be the tree generated by the algorithm (0) if every interval C contained its *expected* number of data points, $np(C)$ (in which case our splitting rule becomes the negation of (1)). If this were the case, the density estimate extracted from this tree would therefore, on each leaf C, be equal to $f(C)$. We begin by showing that Theorem 2 holds for this function, as stated in the following proposition.

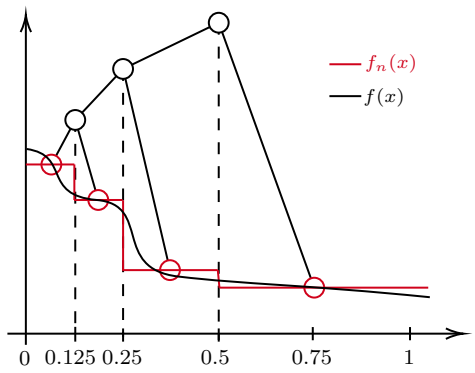

Figure 2. An estimate f_n using a finite, random tree T_n.

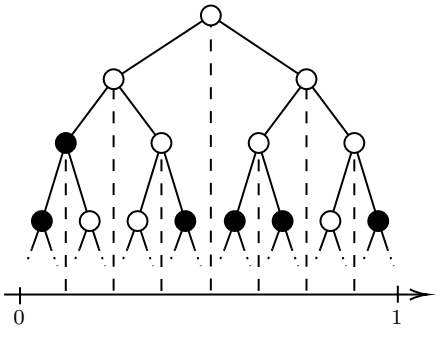

Figure 3. A depiction of an infinite tree $\mathcal{T}_\infty$.

Proposition 5. *Let f be a bounded decreasing probability density function on $[0,1]$ with $B = f(0) < \infty$, and let $\mathcal{F}_n$ be the function that takes the value $f(C)$ on every $C \in \mathcal{P}_0 \cap \mathcal{B}$. Then the L_1 distance between these two functions does not exceed $c_0 \cdot (B^{2/3}/n^{1/3})$ for some constant $c_0 \in \mathbf{R}_{>0}$ that does not depend on B or n.*

6.2. Preliminary results and lemmas

The following three lemmas are needed to prove Proposition 5.

Lemma 6. *For any $C \in \mathcal{T}_\infty$,*

$$p(C_1) - p(C_2) \le \int_C |f - f(C)| \le 2\big(p(C_1) - p(C_2)\big).$$

Proof. Let $x_0 := \sup\{x \in C : f(x) \ge f(C)\}$. Without loss of generality, assume $C = [0,1]$ and $p(C) > 0$. Our result is clear when f is constant on C, so we assume otherwise. Assume first that $x_0 < 1/2$, and define

$$A := \int_0^{x_0} \big(f - f(C)\big), \quad B_1 := \int_{x_0}^{1/2} \big(f(C) - f\big), \quad B_2 := \int_{1/2}^{1} \big(f(C) - f\big).$$

Our assumption on f guarantees that A, B_1 and B_2 are all positive. It is clear that $\int_c |f - f(C)| = A + B_1 + B_2$ and $p(C_1) - p(C_2) = A + (B_2 - B_1)$, which shows the leftmost inequality. Note that $B_2 \ge B_1$, since otherwise we would have

$$p(C_1) - p(C_2) = A + (B_2 - B_1) < A,$$

which would only be possible $|x_0 - 1/2| \ge 1/2$, forcing $x_0 = 0$, $A = 0$ and f to be constant. Using the fact that $A = B_1 + B_2$ (by definition of x_0),

$$2\big(p(C_1) - p(C_2)\big) = A + B_1 + B_2 + 2(B_2 - B_1) \ge \int_C |f - f(C)|.$$

The case $x_0 > 1/2$ can be taken care of similarly. ∎

Lemma 7. *Let $\ell \in \mathbf{Z}^+$ be fixed, and let $\mathcal{A}_\ell$ be the set of nodes in $\mathcal{T}_\infty$ of depth ℓ. Then*

$$\sum_{C \in \mathcal{A}_\ell} \big(p(C_1) - p(C_2)\big) \leq \frac{B}{2^{\ell+1}}.$$

Proof. Let $\{C_i\}_{i=1}^{2^\ell}$ be an enumeration of $\mathcal{A}_\ell$ from left to right (where the left-most node has 0 as one of its interval endpoints). We have

$$\begin{aligned}\sum_{C \in \mathcal{A}_\ell} \big(p(C_1) - p(C_2)\big) &= \sum_{i=1}^{2^\ell} \big(p(C_i') - p(C_i'')\big) \\ &\leq p(C_1') - p(C_{2^\ell}'') \leq p(C_1') \leq \frac{B}{2^{\ell+1}}.\end{aligned}$$

■

Lemma 8. *Let $\ell \in \mathbf{Z}^+$ be fixed, and let $\mathcal{A}_\ell$ be the set of nodes in $\mathcal{T}_\infty$ of depth ℓ. Then*

$$\sum_{C \in \mathcal{A}_\ell} \sqrt{\frac{p(C)}{n}} \leq 2^{\ell/2} \sqrt{\frac{B}{n}}.$$

Proof. By Jensen's inequality, $\sqrt{f(C)} \leq \int_C \sqrt{f}/\lambda(C)$ and

$$\sqrt{p(C)} = \sqrt{\lambda(C) f(C)} \leq \frac{\int_C \sqrt{f}}{\sqrt{\lambda(C)}}.$$

It follows that

$$\sum_{C \in \mathcal{A}_\ell} \sqrt{\frac{p(C)}{n}} \leq 2^{\ell/2} \frac{1}{\sqrt{n}} \int_C \sqrt{f} \leq 2^{\ell/2} \sqrt{\frac{B}{n}}.$$

■

6.3. Proof of proposition 5

Armed with these lemmas, we may now prove Proposition 5.

Proof. The L_1 distance between f and f_n on the whole of $[0,1]$ can be computed by summing the error over the leaf set $\mathcal{B} \cap \mathcal{P}_0$, and is thus equal to

$$\sum_{C \in \mathcal{B} \cap \mathcal{P}_0} \int_C |f - f(C)|.$$

Using Lemma 6 and the definition $\mathcal{B} \cap \mathcal{P}_0$, we can upper bound this quantity and write

$$\sum_{C \in \mathcal{B} \cap \mathcal{P}_0} \int_C |f - f(C)| \leq 2 \cdot \sum_{C \in \mathcal{B} \cap \mathcal{P}_0} \big(p(C_1) - p(C_2)\big) \leq 2 \cdot \sum_{C \in \mathcal{B} \cap \mathcal{P}_0} \gamma \sqrt{\frac{p(C)}{n}}. \tag{2}$$

By Lemmas 7 and 8,

$$\sum_{C\in\mathcal{B}\cap\mathcal{P}_0}\int_C |f-f(C)| \le 2\cdot\sum_{\ell=0}^{\infty}\min\left(\frac{B}{2^{\ell+1}}, \gamma\cdot 2^{\ell/2}\sqrt{\frac{B}{n}}\right). \tag{3}$$

Recall that $\ell^* = \min\{\ell \in \mathbf{Z}_{>0} : B/2^{\ell+1} \le \gamma\cdot 2^{\ell/2}\sqrt{B/n}\}$ and note that ℓ^* is within 1 of

$$\log_2\left\{\left(\frac{Bn}{4}\right)^{1/3}\left(\frac{1}{\gamma}\right)^{2/3}\right\},$$

and that the summation in (3) is bounded above by

$$\begin{aligned}2\left(\sum_{\ell=0}^{\ell^*-1}\gamma\sqrt{\frac{B}{n}}2^{\ell/2} + \sum_{\ell=\ell^*}^{\infty}\frac{B}{2^{\ell+1}}\right) &\le 2\gamma\sqrt{\frac{B}{n}}2^{(\ell^*-1)/2}\left(\frac{1}{1-1/\sqrt{2}}\right) + \frac{2B}{2^{\ell^*}}\\ &\le \frac{\gamma^{2/3}B^{2/3}}{n^{1/3}}\left(\frac{2^{7/6}}{(\sqrt{2}-1)} + 2^{5/3}\right).\end{aligned}$$

This non-asymptotic bound is uniform over all bounded monotone densities f.

■

6.4. Additional results regarding the infinite tree

We conclude this section by stating a few properties of $\mathcal{T}_\infty$ in the following lemmas, which are proved in the appendix. The first is a deterministic bound on the number of unbalanced nodes, both at a given level $\ell \ge \ell^*$ (equation (4)) and in general (equations (5), (6)). The second is a bound on the number of nodes in $\mathcal{T}_\infty$ with exactly j balanced ancestors for a given positive integer j.

Lemma 9. *Let $\mathcal{A}_\ell$ denote the set of nodes in $\mathcal{T}_\infty$ of depth ℓ. If $\ell \ge \ell^*$,*

$$|\mathcal{A}_\ell \setminus \mathcal{B}| \le \frac{2\sqrt{2}}{\gamma}\frac{\sqrt{Bn}}{2^{\ell/2}}, \tag{4}$$

and

$$|\mathcal{B}^c| \le \frac{5B^{1/3}n^{1/3}}{\gamma^{2/3}}. \tag{5}$$

Furthermore,

$$\sup_{\alpha>0}\alpha\left|\left(\mathcal{B}^{(\alpha)}\right)^c\right| \le \frac{5B^{1/3}n^{1/3}}{\gamma^{2/3}}. \tag{6}$$

Lemma 10. *Recall that $\mathcal{B}$ is the subset of balanced nodes and that $\mathcal{B}^c$ is the subset of unbalanced nodes of $\mathcal{T}_\infty$. For any $j \in \mathbf{Z}_{>0}$, define $\mathcal{P}_j = \mathcal{P}_j(\mathcal{T}_\infty)$ to be the set of nodes of $\mathcal{T}_\infty$ with exactly j ancestors in $\mathcal{B}$, then*

$$|\mathcal{P}_j| \le (|\mathcal{B}^c| + 1)\cdot 2^j.$$

7. Proof of Theorem 2

Using the results above, we return to the proof of Theorem 2. The *expected* L_1 distance between f and f_n (as defined previously) is computed by summing over T_n's leaf set, denoted by L. By Scheffé's identity (see Devroye and Györfi, 1985 [8]), we have

$$\mathbf{E}\left\{\int_0^1 |f - f_n|\right\} = 2 \cdot \mathbf{E}\left\{\int_0^1 (f - f_n)_+\right\} \tag{7}$$

where $(x)_+ := \max(x, 0)$. Now, (7) is bounded from above by

$$\underbrace{2\,\mathbf{E}\left\{\sum_{C\in L}\int_C (f - f(C))_+\right\}}_{\text{(I)}} + \underbrace{2\,\mathbf{E}\left\{\sum_{C\in L}\int_C \left(f(C) - \frac{N(C)/n}{\lambda(C)}\right)_+\right\}}_{\text{(II)}}.$$

Here $N(C)$ is the number of data points in C. We bound each of these terms separately.

We view T_n as a sub-tree of $\mathcal{T}_\infty$. This allows us to recycle most of the notation introduced above. For instance, leaves of T_n with depth ℓ are the elements of $L \cap \mathcal{A}_\ell$, while leaves that are balanced are elements of $L \cap \mathcal{B}$.

7.1. Upper bound for (I)

We begin with a few preliminary results.

Lemma 11. *Let C be any non-leaf node of T_n with depth ℓ and let $D \subseteq L$ be the set of leaves of the sub-tree rooted at C, then*

$$\sum_{C^*\in D}\int_{C^*} \big(f - f(C^*)\big)_+ \le \int_C \big(f - f(C)\big)_+.$$

Proof. See appendix. ■

Lemma 12. *Let $C \in T_n \setminus \mathcal{B}^{(\sqrt{2})}$, and $\xi(C) := p(C_1) - p(C_2) - \gamma\sqrt{2p(C)/n} > 0$. Then for such C, we have*

$$\mathbf{P}\{C \in L\} \le \frac{2p(C)}{2p(C) + n\xi(C)^2} + \frac{4}{np(C)}.$$

Proof. See appendix. ■

Using these lemmas, we prove the following proposition.

Proposition 13.

$$\sup_{f\in\mathcal{M}_B} \mathbf{E}\left\{\sum_{C\in L}\int_C (f - f(C))_+\right\} \le \frac{B^{2/3}}{n^{1/3}} c_1(\gamma) + o(n^{-1/3})$$

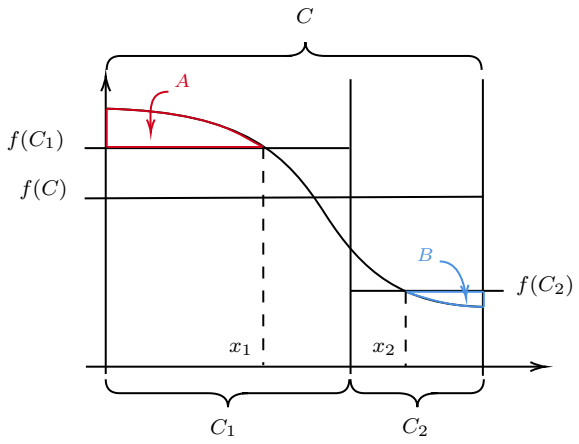

Figure 4. Definitions used in the proof of Lemma 11.

where

$$c_1(\gamma) := \left(4\gamma^{2/3} + \frac{2\sqrt{2}(\gamma + \sqrt{\gamma^2+1})}{\gamma^{1/3}}\right)$$

is a strictly positive constant depending only upon γ.

Proof. The term we are trying to bound can be viewed as the expected L_1 distance between f and the estimator obtained by taking the (random) partition of $[0,1]$ given by T_n, and estimating f by its average value on each interval in the said partition. Informally, one notices that if the branching process that generated T_n behaved "as expected", this estimator would be more or less equal to $\mathcal{F}_n$ from Proposition 5.

Deeper leaves in T_n yield a finer partition of $[0,1]$. Taking intuition from the Riemann integral, one would guess that since we approximate f by its average value on each interval of this partition, a finer partition would help us minimize L_1 distance. Conversely, we can use a coarser partition to upper bound said distance, as shown by Lemma 11.

Thus, we can use the partition given by T_n truncated below level ℓ^* to derive our upper bound. By Lemmas 6 and 11, we have

$$\text{(I)} \leq \mathbf{E}\left\{\sum_{\ell=0}^{\ell^*}\sum_{C\in L\cap\mathcal{A}_\ell}\int_C \big(f - f(C)\big)_+\right\} + 2\cdot\mathbf{E}\left\{\sum_{C\in A_{\ell^*}}\big(p(C_1)-p(C_2)\big)\right\}, \quad (8)$$

and an application of Lemma 7 yields

$$\mathbf{E}\left\{\sum_{C\in A_{\ell^*}}\big(p(C_1)-p(C_2)\big)\right\} \leq \frac{B}{2^{\ell^*+1}} \leq 2\frac{B^{2/3}\gamma^{2/3}}{n^{1/3}}. \quad (9)$$

Next, we recall that $\mathcal{B}^{(\sqrt{2})}$ is the set of nodes of $\mathcal{T}_\infty$ satisfying

$$p(C_1) - p(C_2) \le \gamma\sqrt{\frac{2p(C)}{n}},$$

as defined earlier. Any node C belonging to the complement of $\mathcal{B}^{(\sqrt{2})}$ satisfies

$$p(C_1) - p(C_2) = \gamma\sqrt{\frac{2p(C)}{n}} + \xi(C)$$

where $\xi(C) := p(C_1) - p(C_2) - \gamma\sqrt{2p(C)/n}$ is a strictly positive real number.

We use Lemma 6 once more to bound the leftmost term in (8), writing

$$\begin{aligned}
&\mathbf{E}\biggl\{\sum_{\ell=0}^{\ell^*}\sum_{C\in L\cap\mathcal{A}_\ell}\int_C \bigl(f - f(C)\bigr)_+\biggr\}\\
&\quad= \mathbf{E}\biggl\{\sum_{\ell=0}^{\ell^*}\sum_{C\in\mathcal{A}_\ell}\int_C \bigl(f - f(C)\bigr)_+\mathbf{1}_{[C\in L]}\biggr\}\\
&\quad\le \sum_{\ell=0}^{\ell^*}\biggl(\sum_{C\in\mathcal{B}^{(\sqrt{2})}\cap\mathcal{A}_\ell}\gamma\sqrt{\frac{2p(C)}{n}}\\
&\qquad\qquad + \sum_{C\in\mathcal{A}_\ell\setminus\mathcal{B}^{(\sqrt{2})}}\min\Bigl(\gamma\sqrt{\frac{2p(C)}{n}} + \xi(C),\, p(C_1) - p(C_2)\Bigr)\mathbf{P}\{C\in L\}\biggr).
\end{aligned} \tag{10}$$

Applying Lemma 8, we find that the first of the two inner summations in (10) is bounded above by

$$\gamma\sqrt{\frac{B}{n}}2^{(\ell+1)/2}. \tag{11}$$

To bound the second summation, we use Lemma 12 as well as the fact that $p(C_1) - p(C_2) \le p(C)$ to write

$$\begin{aligned}
&\sum_{C\in\mathcal{A}_\ell\setminus\mathcal{B}^{(\sqrt{2})}}\min\left(\gamma\sqrt{\frac{2p(C)}{n}} + \xi(C),\, p(C_1) - p(C_2)\right)\mathbf{P}\{C\in L\}\\
&\qquad\le \sum_{C\in\mathcal{A}_\ell\setminus\mathcal{B}^{(\sqrt{2})}}\left(\Bigl(\gamma\sqrt{\frac{2p(C)}{n}} + \xi(C)\Bigr)\frac{1}{1 + \xi(C)^2/(2p(C)/n)} + \frac{4}{n}\right).
\end{aligned} \tag{12}$$

For any positive real numbers a and b, the following identity holds:

$$\frac{a+b}{1+b^2} \le \sqrt{a^2+1}.$$

Using it with $a = \gamma$ and $b = \xi(C)/\sqrt{2p(C)/n}$ inside the summation in (12), we have

$$\sum_{C\in\mathcal{A}_\ell\setminus\mathcal{B}^{(\sqrt{2})}}\left(\sqrt{\frac{2p(C)}{n}}\Bigl(\frac{a+b}{1+b^2}\Bigr) + \frac{4}{n}\right) \le \sum_{C\in\mathcal{A}_\ell\setminus\mathcal{B}^{(\sqrt{2})}}\left(\sqrt{\frac{2p(C)}{n}}\sqrt{\gamma^2+1} + \frac{4}{n}\right). \tag{13}$$

Since we are only bounding the quantity above for values of ℓ that are smaller than ℓ^*, we have

$$|\mathcal{A}_\ell \setminus \mathcal{B}^{(\sqrt{2})}| \le |\mathcal{A}_\ell| \le |A_{\ell^*}| \le 2^{\ell^*}.$$

By definition of ℓ^*, this yields

$$|\mathcal{A}_\ell \setminus \mathcal{B}^{(\sqrt{2})}| \le 2^{\ell^*} \le \frac{2(Bn)^{1/3}}{\gamma^{2/3}}.$$

Lemma 8 implies that for any $\ell \le \ell^*$,

$$\sum_{C \in \mathcal{A}_\ell \setminus \mathcal{B}^{(\sqrt{2})}} \left(\sqrt{\frac{2p(C)}{n}} \sqrt{\gamma^2+1} + \frac{4}{n} \right) \le \left(\sqrt{\gamma^2+1} \right) \sqrt{\frac{B}{n}} 2^{(\ell+1)/2} + \frac{8B^{1/3}}{(\gamma n)^{2/3}}.$$

Invoking equations (9) and (11), our bound on (I) in (8) becomes

$$\begin{aligned} \text{(I)} &\le \frac{4B^{2/3}\gamma^{2/3}}{n^{1/3}} + \sum_{\ell=0}^{\ell^*} \left(\left(\gamma + \sqrt{\gamma^2+1} \right) \sqrt{\frac{B}{n}} 2^{(\ell+1)/2} + \frac{8B^{1/3}}{(\gamma n)^{2/3}} \right) \\ &\le \frac{B^{2/3}}{n^{1/3}} c_1(\gamma) + \frac{8(\ell^*+1)B^{1/3}}{(\gamma n)^{2/3}}. \end{aligned}$$

So, $\ell^*/n^{2/3} = O(\log_2(n)/n^{2/3})$ uniformly over all monotone densities bounded by B. This completes the proof of Proposition 13. ■

7.2. Upper bound for (II)

Our upper bound for (II) is given in Proposition 15 below. Its proof relies on the following preliminary result.

Lemma 14. *Assume that $\gamma > 1$. Then for any $0 < \alpha < 1 - 1/\gamma$ and $C \in \mathcal{B}^{(\alpha)} \cap \mathcal{P}_j$, we have*

$$\mathbf{E}\left\{ \left(p(C) - \frac{N(C)}{n} \right)_+ \mathbf{1}_{[C \in L]} \right\} \le c_2(\gamma,\alpha)^{j/2} \sqrt{\frac{p(C)}{n}},$$

where $c_2(\gamma,\alpha) := 1/(1 + (\gamma(1-\alpha))^2) < 1/2$.

Proof. See appendix. ■

Proposition 15. *Assume that $\gamma > 1$. Then we have*

$$\mathbf{E}\left\{ \sum_{C \in L} \int_C \left(f(C) - \frac{N(C)/n}{\lambda(C)} \right)_+ \right\} \le c_3(\gamma) \frac{B^{1/6}}{n^{1/3}} + o(n^{-1/3}),$$

where

$$c_3(\gamma) = \inf_{0<\alpha<1-1/\gamma} \gamma^{-1/3} \sqrt{5} \cdot \left(\frac{1}{\sqrt{\alpha}} + \frac{1}{1 - \sqrt{2c_2(\gamma,\alpha)}} \right).$$

Proof. We start by writing

$$\mathbf{E}\Big\{ \sum_{C\in L} \int_C \Big(f(C) - \frac{N(C)/n}{\lambda(C)} \Big)_+ \Big\} = \mathbf{E}\Big\{ \sum_{C\in L} \Big(p(C) - \frac{N(C)}{n} \Big)_+ \Big\}. \tag{14}$$

Let $0 < \alpha < 1 - 1/\gamma$ be arbitrary. We partition nodes $C \in L$ according to which $\mathcal{P}_j$ they belong to, as well as whether or not they belong to $\mathcal{B}^{(\alpha)}$, seeing that lemmas 9 and 10 provide upper bounds to the number of elements in these sets. We write

$$\begin{aligned}
&\mathbf{E}\Big\{ \sum_{C\in L} \Big(p(C) - \frac{N(C)}{n} \Big)_+ \Big\} \\
&\quad \le \sum_{j=0}^{\infty} \sum_{C\in\mathcal{B}^{(\alpha)}\cap\mathcal{P}_j} \mathbf{E}\Big\{ \Big(p(C) - \frac{N(C)}{n} \Big)_+ \mathbf{1}_{[C\in L]} \Big\} + \sum_{C\notin\mathcal{B}^{(\alpha)}} \sqrt{\frac{p(C)}{n}},
\end{aligned} \tag{15}$$

and then use the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}
\sum_{C\notin\mathcal{B}^{(\alpha)}} \sqrt{\frac{p(C)}{n}} &\le \sqrt{\Big(\sum_{C\notin\mathcal{B}^{(\alpha)}} 1 \Big) \Big(\sum_{C\notin\mathcal{B}^{(\alpha)}} \frac{p(C)}{n} \Big)} \\
&\le \frac{1}{\sqrt{n}} \sqrt{\big|\big(\mathcal{B}^{(\alpha)}\big)^c\big| \cdot \sum_{C\notin\mathcal{B}^{(\alpha)}} p(C)} \\
&\le \frac{1}{\sqrt{n}} \sqrt{\big|\big(\mathcal{B}^{(\alpha)}\big)^c\big|}.
\end{aligned}$$

By Lemma 9, the latter is dominated by

$$\sqrt{\frac{5}{\alpha}} \cdot \frac{B^{1/6}}{\gamma^{1/3} n^{1/3}}.$$

Further, we can use Lemma 14 to write

$$\begin{aligned}
&\sum_{j=0}^{\infty} \sum_{C\in\mathcal{B}^{(\alpha)}\cap\mathcal{P}_j} \mathbf{E}\Big\{ \Big(p(C) - \frac{N(C)}{n} \Big)_+ \mathbf{1}_{[C\in L]} \Big\} \\
&\qquad \le \sum_{j=0}^{\infty} \sum_{C\in\mathcal{B}^{(\alpha)}\cap\mathcal{P}_j} c_2(\gamma,\alpha)^{j/2} \sqrt{\frac{p(C)}{n}} \\
&\qquad \le \frac{1}{\sqrt{n}} \sum_{j=0}^{\infty} c_2(\gamma,\alpha)^{j/2} \sqrt{|\mathcal{P}_j| \cdot \sum_{C\in\mathcal{P}_j} p(C)} \\
&\qquad\qquad \text{(by Jensen's inequality)} \\
&\qquad \le \frac{1}{\sqrt{n}} \sum_{j=0}^{\infty} c_2(\gamma,\alpha)^{j/2} \sqrt{(5\cdot\gamma^{-2/3} B^{1/3} n^{1/3} + 1)\cdot 2^j} \\
&\qquad\qquad \text{(by Lemmas 9 and 10)} \\
&\qquad \le \sqrt{\frac{5\cdot B^{1/3} n^{1/3} + 1}{\gamma^{2/3} n}} \cdot \frac{1}{1 - \sqrt{2c_2(\gamma,\alpha)}} \\
&\qquad = \frac{\sqrt{5}}{1 - \sqrt{2c_2(\gamma,\alpha)}} \cdot \frac{B^{1/6}}{\gamma^{1/3} n^{1/3}} + o(n^{-1/3}),
\end{aligned}$$

and the claim follows since α was picked arbitrarily in $(0, 1 - 1/\gamma)$. ■

Theorem 2 is a direct consequence of propositions 13 and 15.

8. Conclusion

Within the same framework, we can replace the histogram on each set of the partition by a linear estimate with some parameters (slope and intercept at the center point of an interval, for example) only depending upon N_1, N_2, and $\lambda(C)$. Such estimates should adapt better to the smoothness of the density and should be studied for the larger class of bounded monotone densities with bounded first derivative.

Estimates of complexity $\kappa > 2$ could lead to nice and simple estimates for convex, concave, log-convex and log-concave densities. For a concave density, for instance, we sketch how one could decide to split a fixed interval C. Consider four equal–sized sub-intervals $C_i, 1 \leq i \leq 4$, of C, and let N_i be the cardinality of C_i. If C is not split, we estimate f on C by a linear segment with a slope proportional to $(N_3 + N_4) - (N_1 + N_2)$. If the true density were linear on C, then $(N_2 + N_3) - (N_1 + N_4)$ would be of stochastic order $\sqrt{\sum_i N_i}$, for otherwise it would be positively biased. So, a natural splitting rule would be to split C if

$$(N_2 + N_3) - (N_1 + N_4) > \gamma \sqrt{\sum_i N_i}$$

for a fixed design parameter γ.

Finally, one can easily picture extensions to $[0, 1]^d$ for monotone densities (e.g., monotone in each coordinate when all others are fixed). Splitting decisions would then depend upon the 2^d cardinalities of all equal quadrants that partition a cell C. The splits can be binary (along a preferred dimension) or 2^d-ary. In the latter case, one would obtain random quadtrees.

9. Appendix A1: proof of Theorem 1

Following Devroye and Györfi (1985, [8]) and Devroye (1987, [7]), it suffices to show that for all Lebesgue points x with $f(x) > 0$ that $f_n(x) \to f(x)$ in probability. Here we use the fact that almost all x on $[0, 1]$ are Lebesgue points with $f(x) > 0$ (Wheeden and Zygmund, 1977 [16], 2015 [17]), and recall that x is a Lebesgue point for f if

$$\lim_{r \downarrow 0} \sup_{y : x-r \leq y \leq x+r} \frac{1}{r} \int_y^{y+r} f = f(x).$$

We fix such an irrational Lebesgue point x in $(0, 1)$, and introduce the notation $C_0, C_1, \ldots$ for the intervals containing x at levels $0, 1, \ldots$ in the binary tree.

Thus, $C_0 = [0,1]$, C_1 is either $[0,1/2]$ or $[1/2,1]$, and so forth. Let $N_i = N(C_i)$ be the cardinality of interval C_i. Let K be the level at which we find the first leaf on the path to x in the binary tree. Also, let $U_0, U_1, \ldots$ be the sequence of uniform random variables used for the randomized splitting at each level. In other words, C_K is the first un-split interval, i.e., the sole leaf interval on that path. We first show that $K \to \infty$ and $n/2^K \to \infty$ in probability as $n \to \infty$.

Note that for any large but fixed integer k, we have for any integer m,

$$\begin{aligned}\mathbf{P}\{K \le k\} &\le \mathbf{E}\left\{\sum_{i=0}^{k} \varphi(N_i)\right\} \\ &\le (k+1)\mathbf{E}\left\{\varphi(N_k)\right\} \\ &\le (k+1)\mathbf{P}\{N_k \le m\} + (k+1)\varphi(m).\end{aligned}$$

We can pick m large enough to make the last term as small as desired. Since N_k is binomial (n, p_k), where $p_k = \int_{C_k} f$, we have $\mathbf{P}\{N_k \le m\} = o(1)$. Therefore, $K \to \infty$ in probability.

Next, for any large but fixed integer k, we have for any positive integer m

$$\begin{aligned}\mathbf{P}\{n/2^K \le k\} &= \mathbf{P}\{K \ge \log_2(n/k)\} \\ &\le \mathbf{E}\left\{\prod_{i<\log_2(n/k)} (1-\varphi(N_i))\right\} \\ &\le \mathbf{E}\left\{\exp\left(-\sum_{i<\log_2(n/k)} \varphi(N_i)\right)\right\} \\ &\le \mathbf{E}\left\{\exp\left(-\sum_{\log_2(n/(km))\le i<\log_2(n/k)} \varphi(N_i)\right)\right\} \\ &\le \mathbf{E}\left\{\exp\left(-\log_2(m)\,\varphi\left(N_{\lfloor \log_2(n/(km))\rfloor}\right)\right)\right\}.\end{aligned}$$

Now,

$$\varphi\left(N_i\right) \ge \varphi(\ell)$$

if $N_i \le \ell$, where $i = \lfloor \log_2(n/(km))\rfloor$. Thus,

$$\mathbf{P}\{n/2^K \le k\} \le \mathbf{P}\{N_i > \ell\} + e^{-\log_2 m \times \varphi(\ell)}.$$

By the Lebesgue density theorem, $2^i \int_{C_i} f \to f(x)$ as n (and thus i) tends to ∞. Therefore, there exists a finite constant c such that $\sup_i \int_{C_i} f \le c/2^i$, and thus,

$$\mathbf{E}\{N_i\} = n\int_{C_i} f \le \frac{cn}{2^i} \le 2ckm.$$

By Markov's inequality,

$$\begin{aligned}\mathbf{P}\{n/2^K \le k\} &\le \frac{\mathbf{E}\{N_i\}}{\ell} + e^{-\log_2 m \times \varphi(\ell)} \\ &\le \frac{2ckm}{\ell} + e^{-\log_2 m \times \varphi(\ell)}.\end{aligned}$$

We take $\ell = m^2$ and pick m large enough to make the first term small. Since $\log_2(m^2)\varphi(m^2) \to \infty$, the second term can also be made as small as desired by picking m large enough. We conclude that $n/2^K \to \infty$ in probability.

Let us denote the histogram estimate at x based on the i-th level interval C_i by

$$g_i(x) = 2^i \frac{N_i}{n}.$$

For $\epsilon > 0$ and integer k, we have

$$\begin{aligned}&\mathbf{P}\left\{|f_n(x) - f(x)| > \epsilon\right\} \\ &\quad \le \mathbf{P}\{K \le k\} + \mathbf{P}\{K \ge \log_2(n) - k\} \\ &\qquad + \mathbf{P}\left\{\cup_{i=k}^{\log_2(n)-k} \left[|g_i(x) - f(x)| > \epsilon\right]\right\}.\end{aligned}$$

By choice of k, the first term can be made as small as desired, while the second term is $o(1)$. The third term is controlled by the union bound,

$$\sum_{i=k}^{\log_2(n)-k} \mathbf{P}\{|g_i(x) - f(x)| > \epsilon\}.$$

Note that

$$|\mathbf{E}\{g_i(x)\} - f(x)| = \left|2^i \int_{C_i} f - f(x)\right| \le \frac{\epsilon}{2}$$

when k (and thus i) is large enough. Using $\mathbf{V}$ to denote the variance, we have

$$\begin{aligned}\mathbf{V}\{g_i(x)\} &= \frac{2^{2i}}{n^2}\mathbf{V}\{N_i\} \\ &\le \frac{2^{2i}\int_{C_i} f}{n} \\ &\le \frac{2^i(f(x) + \epsilon/2)}{n}.\end{aligned}$$

So, by Chebyshev's inequality,

$$\begin{aligned}\sum_{i=k}^{\log_2(n)-k} \mathbf{P}\{|g_i(x) - f(x)| > \epsilon\} &\le \sum_{i=k}^{\log_2(n)-k} \mathbf{P}\{|g_i(x) - \mathbf{E}\{g_i(x)\}| > \epsilon\} \\ &\le \frac{4}{\epsilon^2} \sum_{i=k}^{\log_2(n)-k} \mathbf{V}\{g_i(x)\} \\ &\le \frac{4}{\epsilon^2} \sum_{i=k}^{\log_2(n)-k} \frac{2^i(f(x) + \epsilon/2)}{n} \\ &\le \frac{8(f(x) + \epsilon/2)}{2^k\epsilon^2},\end{aligned}$$

and this is as small as desired by picking k large enough. This concludes the proof of Theorem 1.

10. Appendix A2: proof of Lemma 9

List the unbalanced nodes of $\mathcal{A}_\ell$ in order from right to left, where the leftmost node is that for which the left interval endpoint is the smallest. Denote this list $\{C_i\}_{i=1}^k$, where $k = |\mathcal{A}_\ell \setminus \mathcal{B}|$.

By the monotonicity of f, we have $p(C_0) \le p(C_1) \le \cdots \le p(C_k)$, and we can therefore write $p(C_i) = \sum_{j=0}^i q_j$ for every i, where $q_1, \ldots, q_k$ are non-negative. Since every C_i is unbalanced, we have $p(C_i') - p(C_i'') > \gamma\sqrt{p(C_i)/n}$ which, combined with the fact that $p(C_i) = p(C_i') + p(C_i'')$, yields $2p(C_i') \ge \gamma\sqrt{p(C_i)/n} + p(C_i)$ and in turn

$$p(C_{i+1}) \ge \gamma\sqrt{\frac{p(C_i)}{n}} + p(C_i)$$

for any $1 \le i \le k$. We use this fact to prove that for any $1 \le i \le k$, $q_i \ge (\gamma^2/4n)(i+1)$. If $i = 0$, we have $q_0 = p(C_0) \ge \gamma^2/n \ge \gamma^2/(4n)$. Now assume that the claim regarding q_i holds for some i, then

$$\begin{aligned} q_{i+1} &= p(C_{i+1}) - p(C_i) \\ &\ge \gamma\sqrt{\frac{p(C_i)}{n}} \\ &\ge \frac{\gamma}{\sqrt{n}}\Big(\frac{\gamma^2}{4n}\frac{(i+1)(i+2)}{2}\Big)^{1/2} \\ &\ge \frac{\gamma^2}{4n}(i+2) \end{aligned}$$

and the claim follows by induction. Therefore,

$$p(C_k) = \sum_{i=1}^k q_k \ge \frac{\gamma^2}{4n}\sum_{i=1}^k (i+1) \ge \frac{\gamma^2 k^2}{8n},$$

and the first part of the lemma follows since $p(C_k) \le B/2^\ell$. The upper bound on $|\mathcal{B}^c|$ follows from the fact that it is no larger than $2^{\ell^*} + \sum_{\ell \ge \ell^*} |\mathcal{A}_\ell \setminus \mathcal{B}|$. Lastly, an identical argument yields the upper bound for $|(\mathcal{B}^{(\alpha)})^c|$.

11. Appendix A3: proof of Lemma 10

We begin by noticing that all but finitely many nodes of $\mathcal{T}_\infty$ are in $\mathcal{B}$. It follows that for any $i \in \mathbf{Z}_{>0}$, $|\mathcal{P}_i| \le |\mathcal{P}_{i+1}|$ since any node in $\mathcal{P}_i$ is the root of a tree that contains at least one balanced node in $\mathcal{P}_{i+1}$.

Next, we examine how switching a balanced node with its parent affects the various $|\mathcal{P}_i|$'s. Let C be an arbitrary balanced node of $\mathcal{T}_\infty$, D be its parent

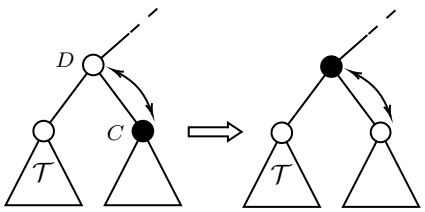

Figure 5. Switching a balanced node with its parent.

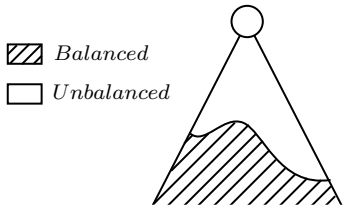

Figure 6. Tree for which $|\mathcal{P}_j|$ is maximized (for all j).

and a be the number of balanced ancestors of D. We may assume that D is unbalanced since switching D and C would leave the tree unaffected. Our operation is depicted in figure 5. If we let $\mathcal{T}$ be the sub-tree of $\mathcal{T}_\infty$ rooted at C's sibling, then switching C and D applies the map

$$|\mathcal{P}_j| \mapsto \begin{cases} |\mathcal{P}_j| & \text{if } j \geq a \\ |\mathcal{P}_j| - |\mathcal{P}_j(\mathcal{T})| + |\mathcal{P}_{j+1}(\mathcal{T})| & \text{if } j < a \end{cases}$$

to every $|\mathcal{P}_j|$.

Since $|\mathcal{P}_j(\mathcal{T})| \leq |\mathcal{P}_{j+1}(\mathcal{T})|$ for any j, this map's output is always greater than or equal to $|\mathcal{P}_j|$. In other words, the node configuration that maximizes $|\mathcal{P}_j|$ for every j is such that all the unbalanced nodes are pushed to the top. This forms a tree of unbalanced nodes whose leaves are the roots of full, infinite binary trees where all nodes are balanced.

In this configuration, it is clear that $|\mathcal{P}_0| \leq (|\mathcal{B}^c|+1|)$ and that $|\mathcal{P}_{j+1}| \leq 2|\mathcal{P}_j|$, from which the lemma follows.

12. Appendix A4: proof of Lemma 11

It suffices to show that for all pairs of disjoint intervals $C_1, C_2 \subseteq C$ such that $C_1 \cup C_2 = C$

$$\int_{C_1} \big(f - f(C_1)\big)_+ + \int_{C_2} \big(f - f(C_2)\big)_+ \leq \int_C \big(f - f(C)\big)_+.$$

The lemma then follows by induction. Without loss of generality, assume that $C = [0, 1]$.

Let C_1, C_2 be an arbitrary such pair and note that $f(C_2) \leq f(C) \leq f(C_1)$ since f is decreasing. Let $x_1 := \sup\{x : f(x) \geq f(C_1)\}$, $x_2 := \sup\{x : f(x) \geq f(C_2)\}$, and let $A := \int_0^{x_1}(f - f(C_1))$ and $B := \int_{x_2}^1(f - f(C_2))$ as depicted in figure 4. Then it is obvious that

$$A + B = \int_{C_1} \big(f - f(C_1)\big)_+ + \int_{C_2} \big(f - f(C_2)\big)_+ \leq \int_C \big(f - f(C)\big)_+.$$

13. Appendix A5: proof of Lemma 12

Let $C \in T_n$ be arbitrary. Recall that $N(C)$ is the number of data points lying in C (similarly, $N(C_1)$ and $N(C_2)$ are the number of points lying in the left/right halves of C respectively). Notice that $N(C) \stackrel{\mathcal{L}}{=} \mathrm{Bin}(n, p(C))$ (where $\stackrel{\mathcal{L}}{=}$ denotes equivalence in law and $\mathrm{Bin}(n,p)$ a binomial n, p). For C to be a leaf, we must have decided not to split its node. Therefore

$$\begin{aligned}\mathbf{P}\{C \in L\} &\leq \mathbf{P}\Big\{N(C_1) - N(C_2) < \gamma\sqrt{N(C)}\Big\} \\ &\leq \sup_{m \geq np(C)/2} \mathbf{P}\Big\{N(C_1) - N(C_2) < \gamma\sqrt{N(C)} \;\Big|\; N(C) = m\Big\} \\ &\quad + \mathbf{P}\big\{|N(C) - np(C)| > np(C)/2\big\}. \end{aligned} \tag{16}$$

A simple application of Chebyshev's inequality gives

$$\mathbf{P}\big\{|N(C) - np(C)| > np(C)/2\big\} \leq \frac{4(1-p(C))}{np(C)} \leq \frac{4}{np(C)}.$$

Next, given $N(C) = m$, $N(C_1) \stackrel{\mathcal{L}}{=} \mathrm{Bin}(m, p(C_1)/p(C))$. Note that

$$\begin{aligned} p(C_1) + p(C_2) &= p(C) \\ p(C_1) - p(C_2) &= \gamma\sqrt{\frac{2p(C)}{n}} + \xi(C) \end{aligned}$$

so that

$$\frac{p(C_1)}{p(C)} = \frac{1}{2} + \frac{1}{2}\gamma\sqrt{\frac{2}{np(C)}} + \frac{1}{2}\frac{\xi(C)}{p(C)}.$$

Using these observations and the fact that $N(C_1) - N(C_2) = 2N(C_1) - N(C)$, we find that

$$\begin{aligned} &\mathbf{P}\Big\{N(C_1) - N(C_2) < \gamma\sqrt{N(C)} \;\Big|\; N(C) = m\Big\} \\ &= \mathbf{P}\left\{\mathrm{Bin}\left(m, \frac{p(C_1)}{p(C)}\right) - m\frac{p(C_1)}{p(C)} < \gamma\frac{\sqrt{m}}{2} - \gamma\frac{m}{2}\sqrt{\frac{2}{np(C)}} - \frac{m}{2}\frac{\xi(C)}{p(C)}\right\} \\ &\leq \mathbf{P}\left\{\mathrm{Bin}\left(m, \frac{p(C_1)}{p(C)}\right) - m\frac{p(C_1)}{p(C)} < -\frac{m}{2}\frac{\xi(C)}{p(C)}\right\} \end{aligned}$$

if $m \geq np(C)/2$, which is the case in the supremum taken in (16). Using the Chebyshev-Cantelli inequality (see Lugosi, Massart and Boucheron, 2013 [10]) and the fact that the variance of a binomial n, p is at most $n/4$, we have

$$\begin{aligned} &\mathbf{P}\left\{\mathrm{Bin}\left(m, \frac{p(C_1)}{p(C)}\right) - m\frac{p(C_1)}{p(C)} < -\frac{m}{2}\frac{\xi(C)}{p(C)}\right\} \\ &\leq \frac{m/4}{m/4 + \frac{m^2\xi(C)^2}{4(p(C))^2}} \\ &\leq \frac{2p(C)}{2p(C) + n\xi(C)^2} \quad \left(\text{since } m \geq \frac{np(C)}{2}\right) \end{aligned}$$

and the lemma follows.

14. Appendix A6: proof of Lemma 14

Using the Cauchy-Schwarz inequality, we write

$$\mathbf{E}\Big\{\mathbf{1}_{[C\in L]}\Big(p(C)-\frac{N(C)}{n}\Big)\Big\} \le \sqrt{\mathbf{P}\{C\in L, N(C)\le np(C)\}}\cdot\sqrt{\frac{p(C)}{n}}.$$

We claim that

$$\mathbf{P}\{C\in L, N(C)\le np(C)\}\le c_2(\gamma,\alpha)^j,$$

where $c_2(\gamma,\alpha)$ is defined in the lemma's statement. This bound does not depend on the level ℓ at which the node is located. To prove this, it suffices to show that for any balanced node $C\in\mathcal{B}^{(\alpha)}$, if $N(C)\le np(C)$,

$$\mathbf{P}\{N(C_1)-N(C_2)>\gamma\sqrt{N(C)}\mid N(C)\}\le c_2(\gamma,\alpha),$$

since consecutive splits are independent given $N(C)$. For simplicity, temporarily denote $N(C)$, $N(C_1)$ and $N(C_2)$ by N, N_1 and N_2 respectively, and similarly for p, p_1, p_2. Then observe that

$$\begin{aligned}
&\mathbf{P}\Big\{N_1-N_2>\gamma\sqrt{N}\;\Big|\;N\Big\}\\
&\quad\le \mathbf{P}\Big\{N_1-N_2-N\Big(\frac{p_1-p_2}{p}\Big)>\gamma\sqrt{N}-N\Big(\frac{p_1-p_2}{p}\Big)\;\Big|\;N\Big\}\\
&\quad\le \mathbf{1}_{[N(p_1-p_2)/p\ge\gamma\alpha\sqrt{N}]}+\mathbf{P}\Big\{N_1-N_2-N\Big(\frac{p_1-p_2}{p}\Big)>\gamma(1-\alpha)\sqrt{N}\;\Big|\;N\Big\}\\
&\quad\le \mathbf{1}_{[\sqrt{N}>(\gamma\alpha)(p/(p_1-p_2))]}+\mathbf{P}\left\{\frac{N_1-N_2-N\cdot\frac{p_1-p_2}{p}}{\sqrt{4N\cdot\frac{p_1}{p}\frac{p_2}{p}}}>\gamma(1-\alpha)\frac{p/2}{\sqrt{p_1p_2}}\;\Big|\;N\right\}.
\end{aligned}$$

By definition of $\mathcal{B}^{(\alpha)}$, we have $p_1-p_2<(\gamma\alpha)\sqrt{p/n}$. If the indicator in the expression above were one, this would imply that $\sqrt{N}>\sqrt{np}$ which cannot be true since $N\le np$, hence the indicator is equal to 0. As for the probability term, notice that

$$\sqrt{4N\cdot\frac{p_1}{p}\frac{p_2}{p}}$$

is the conditional variance of

$$N_1-N_2-N\cdot\frac{p_1-p_2}{p},$$

which is a random variable with mean 0 and unit variance. Noting that $p/2\ge\sqrt{p_1p_2}$ and that $\gamma(1-\alpha)>1$ by assumption, we can apply the Chebyshev-Cantelli inequality to get

$$\mathbf{P}\left\{\frac{N_1-N_2-N\cdot\frac{p_1-p_2}{p}}{\sqrt{4N\cdot\frac{p_1}{p}\frac{p_2}{p}}}>\gamma(1-\alpha)\frac{p/2}{\sqrt{p_1p_2}}\;\Big|\;N\right\}\le\frac{1}{1+(\gamma-\alpha\gamma)^2}\stackrel{\text{def}}{=}c_2(\gamma,\alpha)$$

and $c_1(\gamma,\alpha)<1/2$.

15. Appendix A7: proof of corollary 3

The decision to split an interval containing k points can be made in order $\log_2(k)$ time, which is the time taken to determine which points lie on the left and right halves of the interval, respectively, via binary search. It therefore suffices to show that the expected number of leaves of the tree generated by the algorithm is of the order of $n^{1/3}$. Letting L denote the tree's leaf set and $\alpha = (\gamma - 1)/2\gamma$, we have

$$\begin{aligned}\mathbf{E}\{|L|\} &\leq \mathbf{E}\big\{\big|\big(\mathcal{B}^{(\alpha)}\big)^c\big|\big\} + \mathbf{E}\big\{\big|L \cap \mathcal{B}^{(\alpha)}\big|\big\} \\ &= \mathbf{E}\big\{\big|\big(\mathcal{B}^{(\alpha)}\big)^c\big|\big\} + \sum_{j=0}^{\infty} \sum_{C \in \mathcal{B}^{(\alpha)} \cap \mathcal{P}_j} \mathbf{E}\big\{\mathbf{1}_{[C \in L]}\big\}.\end{aligned}$$

The first term is $O(n^{1/3})$ by Lemma 9 and the second by the proofs of Proposition 15 and Lemma 14.

16. Acknowledgments

The authors would like to thank both referees.

References

[1] G. Alsmeyer. *Galton–Watson Processes.* Course notes at the University of Münster, 2008.

[2] L. Anderlini. Density estimation trees as fast non-parametric modelling tools. *arXiv*, 1607.06635v1, 2016.

[3] K. B. Athreya and P. E. Ney. *Branching Processes.* Die Grundlehren der mathematischen Wissenschaften, Band 196. Springer-Verlag, New York-Heidelberg, 1972.

[4] A. C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49:122–136, 1941.

[5] G. Biau and L. Devroye. Cellular tree classifiers. *Electronic Journal of Statistics*, 7:1875–1912, 2013.

[6] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, pages 181–237, 1983.

[7] L. Devroye. *A Course in Density Estimation.* Boston, Birkhäuser Verlag, 1987.

[8] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L1 View*. Wiley series in probability and mathematical statistics. New York, John Wiley, 1985.

[9] U. Grenander. On the theory of mortality measurement. *Skandinavisk Aktuarietidskrift*, pages 125–153, 1956.

[10] G. Lugosi, P. Massart, and S. Boucheron. *Concentration Inequalities. A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[11] R. Lyons and Y. Peres. *Probability on Trees and Networks*, volume 42 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York, 2016.

[12] V. V. Petrov. *Sums of Independent Random Variables*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer, Berlin, Heidelberg, 1975.

[13] P. Ram and A. G. Gray. Density estimation trees. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA*, pages 627–635, 2011.

[14] G. Schmidberger. *Tree-based Density Estimation: Algorithms and Applications*. PhD thesis, University of Waikato, New Zealand, 2009.

[15] L. Wasserman. *All of Nonparametric Statistics*. Springer New York, NY, 2006.

[16] R. L. Wheeden and A. Zygmund. *Measure and Integral: An Introduction to Real Analysis*. Number 43 in Monographs and Textbooks in Pure and Applied Mathematics. Marcel Dekker, New York, 1977.

[17] R. L. Wheeden and A. Zygmund. *Measure and Integral: An Introduction to Real Analysis*. Number 43 in Monographs and Textbooks in Pure and Applied Mathematics. CRC Press, Boca Raton, FL, 2nd edition, 2015.