A. Algorithm pseudo-code

We present pseudo-code for the basic algorithm only, without the bounded fringe technique described in Section 3.6. The addition of a bounded fringe is straightforward, but complicates the presentation significantly.

Candidate split dimension A dimension along which a split may be made.

Candidate split point One of the first m structure points to arrive in a leaf.

Candidate split A combination of a candidate split dimension and a position along that dimension to split. These are formed by projecting each candidate split point into each candidate split dimension.

Candidate children Each candidate split in a leaf induces two candidate children for that leaf. These are also referred to as the left and right child of that split.

$N^e(A)$ is a count of estimation points in the cell A, and $Y^e(A)$ is the histogram of labels of these points in A. $N^s(A)$ and $Y^s(A)$ are the corresponding values derived from structure points.

Algorithm 1 BuildTree

Require: Initially the tree has exactly one leaf (TreeRoot) which covers the whole space
Require: The dimensionality of the input, D. Parameters λ, m and τ.

```
SelectCandidateSplitDimensions(TreeRoot, min(1 + Poisson(λ), D))
for t = 1 ... do
  Receive (X_t, Y_t, I_t) from the environment
  A_t ← leaf containing X_t
  if I_t = estimation then
    UpdateEstimationStatistics(A_t, (X_t, Y_t))
    for all S ∈ CandidateSplits(A_t) do
      for all A ∈ CandidateChildren(S) do
        if X_t ∈ A then
          UpdateEstimationStatistics(A, (X_t, Y_t))
        end if
      end for
    end for
  else if I_t = structure then
    if A_t has fewer than m candidate split points then
      for all d ∈ CandidateSplitDimensions(A_t) do
        CreateCandidateSplit(A_t, d, π_d X_t)
      end for
    end if
    for all S ∈ CandidateSplits(A_t) do
      for all A ∈ CandidateChildren(S) do
        if X_t ∈ A then
          UpdateStructuralStatistics(A, (X_t, Y_t))
        end if
      end for
    end for
    if CanSplit(A_t) then
      if ShouldSplit(A_t) then
        Split(A_t)
      else if MustSplit(A_t) then
        Split(A_t)
      end if
    end if
  end if
end for
```

Algorithm 2 Split

Require: A leaf A
Require: At least one valid candidate split for exists for A
$S \leftarrow \text{BestSplit}(A)$
$A' \leftarrow \text{LeftChild}(A)$
SelectCandidateSplitDimensions(A', $\min(1 + \text{Poisson}(\lambda), D)$)
$A'' \leftarrow \text{RightChild}(A)$
SelectCandidateSplitDimensions(A'', $\min(1 + \text{Poisson}(\lambda), D)$)
return A', A''

Algorithm 3 CanSplit

Require: A leaf A
$d \leftarrow \text{Depth}(A)$
for all $S \in \text{CandidateSplits}(A)$ **do**
if SplitIsValid(A, S) **then**
return true
end if
end for
return false

Algorithm 4 SplitIsValid

Require: A leaf A
Require: A split S
$d \leftarrow \text{Depth}(A)$
$A' \leftarrow \text{LeftChild}(S)$
$A'' \leftarrow \text{RightChild}(S)$
return $N^e(A') \geq \alpha(d)$ and $N^e(A'') \geq \alpha(d)$

Algorithm 5 MustSplit

Require: A leaf A
$d \leftarrow \text{Depth}(A)$
return $N^e(A) \geq \beta(d)$

Algorithm 6 ShouldSplit

Require: A leaf A
for all $S \in \text{CandidateSplits}(A)$ **do**
if InformationGain(S) $> \tau$ **then**
if SplitIsValid(A, S) **then**
return true
end if
end if
end for
return false

Algorithm 7 BestSplit

Require: A leaf A
Require: At least one valid candidate split exists for A
best_split $\leftarrow$ **none**
for all $S \in \text{CandidateSplits(A)}$ **do**
if InformationGain(A, S) $\geq$ InformationGain(A, **best_split**) **then**
if SplitIsValid(A, S) **then**
best_split $\leftarrow S$
end if
end if
end for
return best_split

Algorithm 8 InformationGain

Require: A leaf A
Require: A split S
$A' \leftarrow \text{LeftChild}(S)$
$A'' \leftarrow \text{RightChild}(S)$
return $\text{Entropy}(Y^s(A)) - \frac{N^s(A')}{N^s(A)} \text{Entropy}(Y^s(A')) - \frac{N^s(A'')}{N^s(A)} \text{Entropy}(Y^s(A''))$

Algorithm 9 UpdateEstimationStatistics

Require: A leaf A
Require: A point (X, Y)
$N^e(A) \leftarrow N^e(A) + 1$
$Y^e(A) \leftarrow Y^e(A) + Y$

Algorithm 10 UpdateStructuralStatistics

Require: A leaf A
Require: A point (X, Y)
$N^s(A) \leftarrow N^s(A) + 1$
$Y^s(A) \leftarrow Y^s(A) + Y$

B. Proof of Consistency

B.1. A note on notation

A will be reserved for subsets of $\mathbb{R}^D$, and unless otherwise indicated it can be assumed that A denotes a cell of the tree partition. We will often be interested in the cell of the tree partition containing a particular point, which we denote $A(x)$. Since the partition changes over time, and therefore the shape of $A(x)$ changes as well, we use a subscript to disambiguate: $A_t(x)$ is the cell of the partition containing x at time t. Cells in the tree partition have a lifetime which begins when they are created as a candidate child to an existing leaf and ends when they are themselves split into two children. When referring to a point $X_\tau \in A_t(x)$ it is understood that τ is restricted to the lifetime of $A_t(x)$.

We treat cells of the tree partition and leafs of the tree defining it interchangeably, denoting both with an appropriately decorated A.

N generally refers to the number of points of some type in some interval of time. A superscript always denotes type, so N^k refers to a count of points of type k. Two special types, e and s, are used to denote estimation and structure points, respectively. Pairs of subscripts are used to denote time intervals, so $N^k_{a,b}$ denotes the number of points of type k which appear during the time interval $[a, b]$. We also use N as a function whose argument is a subset of $\mathbb{R}^D$ in order to restrict the counting spatially: $N^e_{a,b}(A)$ refers to the number of estimation points which fall in the set A during the time interval $[a, b]$. We make use of one additional variant of N as a function when its argument is a cell in the partition: when we write $N^k(A_t(x))$, without subscripts on N, the interval of time we count over is understood to be the lifetime of the cell $A_t(x)$.

B.2. Preliminaries

Lemma 6. *Suppose we partition a stream of data into c parts by assigning each point (X_t, Y_t) to part $I_t \in \{1, \dots, c\}$ with fixed probability p_k, meaning that*

$$N^k_{a,b} = \sum_{t=a}^{b} \mathbb{I}\{I_t = k\} \quad . \tag{1}$$

Then with probability 1, $N^k_{a,b} \to \infty$ for all $k \in \{1, \dots, c\}$ as $b - a \to \infty$.

Proof. Note that $\mathbb{P}(I_t = 1) = p_1$ and these events are independent for each t. By the second Borel-Cantelli lemma, the probability that the events in this sequence occur infinitely often is 1. The cases for $I_t \in \{2, \dots, c\}$ are similar. □

Lemma 7. *Let X_t be a sequence of iid random variables with distribution μ, let A be a fixed set such that $\mu(A) > 0$ and let $\{I_t\}$ be a fixed partitioning sequence. Then the random variable*

$$N^k_{a,b}(A) = \sum_{a \le t \le b : I_t = k} \mathbb{I}\{X_t \in A\}$$

is Binomial with parameters $N^k_{a,b}$ and $\mu(A)$. In particular,

$$\mathbb{P}\left(N^k_{a,b}(A) \le \frac{\mu(A)}{2} N^k_{a,b}\right) \le \exp\left(-\frac{\mu(A)^2}{2} N^k_{a,b}\right)$$

which goes to 0 as $b - a \to \infty$, where $N^k_{a,b}$ is the deterministic quantity defined as in Equation 1.

Proof. $N^k_{a,b}(A)$ is a sum of iid indicator random variables so it is Binomial. It has the specified parameters because it is a sum over $N^k_{a,b}$ elements and $\mathbb{P}(X_t \in A) = \mu(A)$. Moreover, $\mathbb{E}\left[N^k_{a,b}(A)\right] = \mu(A) N^k_{a,b}$ so by Hoeffding's inequality we have that

$$\mathbb{P}\left(N^k_{a,b}(A) \le \mathbb{E}\left[N^k_{a,b}(A)\right] - \epsilon N^k_{a,b}\right) = \mathbb{P}\left(N^k_{a,b}(A) \le N^k_{a,b}(\mu(A) - \epsilon)\right) \le \exp\left(-2\epsilon^2 N^k_{a,b}\right) \quad .$$

Setting $\epsilon = \frac{1}{2}\mu(A)$ gives theresult. □

B.3. Proof of Proposition 2

Proof. Let $g(x)$ denote the Bayes classifier. Consistency of $\{g_t\}$ is equivalent to saying that $\mathbb{E}[L(g_t)] = \mathbb{P}(g_t(X,Z) \neq Y) \to L^*$. In fact, since $\mathbb{P}(g_t(X,Z) \neq Y \mid X = x) \geq \mathbb{P}(g(X) \neq Y \mid X = x)$ for all $x \in \mathbb{R}^D$, consistency of $\{g_t\}$ means that for μ-almost all x,

$$\mathbb{P}(g_t(X,Z) \neq Y \mid X = x) \to \mathbb{P}(g(X) \neq Y \mid X = x) = 1 - \max_k\{\eta^k(x)\}$$

Define the following two sets of indices

$$G = \{k \mid \eta^k(x) = \max_k\{\eta^k(x)\}\} \ ,$$
$$B = \{k \mid \eta^k(x) < \max_k\{\eta^k(x)\}\} \ .$$

Then

$$\begin{aligned}\mathbb{P}(g_t(X,Z) \neq Y \mid X = x) &= \sum_k \mathbb{P}(g_t(X,Z) = k \mid X = x)\,\mathbb{P}(Y \neq k | X = x) \\ &\leq (1 - \max_k\{\eta^k(x)\}) \sum_{k \in G} \mathbb{P}(g_t(X,Z) = k \mid X = x) + \sum_{k \in B} \mathbb{P}(g_t(X,Z) = k \mid X = x) \ ,\end{aligned}$$

which means it suffices to show that $\mathbb{P}\left(g_t^{(M)}(X, Z^M) = k \mid X = x\right) \to 0$ for all $k \in B$. However, using Z^M to denote M (possibly dependent) copies of Z, for all $k \in B$ we have

$$\begin{aligned}\mathbb{P}\left(g_t^{(M)}(x, Z^M) = k\right) &= \mathbb{P}\left(\sum_{j=1}^M \mathbb{I}\{g_t(x, Z_j) = k\} > \max_{c \neq k} \sum_{j=1}^M \mathbb{I}\{g_t(x, Z_j) = c\}\right) \\ &\leq \mathbb{P}\left(\sum_{j=1}^M \mathbb{I}\{g_t(x, Z_j) = k\} \geq 1\right)\end{aligned}$$

By Markov's inequality,

$$\begin{aligned}&\leq \mathbb{E}\left[\sum_{j=1}^M \mathbb{I}\{g_t(x, Z_j) = k\}\right] \\ &= M\mathbb{P}(g_t(x,Z) = k) \to 0\end{aligned}$$

□

B.4. Proof of Proposition 3

Proof. The sequence in question is uniformly integrable, so it is sufficient to show that $\mathbb{E}[\mathbb{P}(g_t(X,Z,I) \neq Y \mid I)] \to L^*$ implies the result, where the expectation is taken over the random selection of training set.

We can write

$$\begin{aligned}\mathbb{P}(g_t(X,Z,I) \neq Y) &= \mathbb{E}[\mathbb{P}(g_t(X,Z,I) \neq Y \mid I)] \\ &= \int_{\mathcal{I}} \mathbb{P}(g_t(X,Z,I) \neq Y \mid I)\,\nu(I) + \int_{\mathcal{I}^c} \mathbb{P}(g_t(X,Z,I) \neq Y \mid I)\,\nu(I)\end{aligned}$$

By assumption $\nu(\mathcal{I}^c) = 0$, so we have

$$\lim_{t\to\infty} \mathbb{P}(g_t(X,Z,I) \neq Y) = \lim_{t\to\infty} \int_{\mathcal{I}} \mathbb{P}(g_t(X,Z,I) \neq Y \mid I)\,\nu(I)$$

Since probabilities are bounded in the interval $[0, 1]$, the dominated convergence theorem allows us to exchange the integral and the limit,

$$= \int_{\mathcal{I}} \lim_{t\to\infty} \mathbb{P}\left(g_t(X, Z, I) \neq Y \mid I\right) \nu(I)$$

and by assumption the conditional risk converges to the Bayes risk for all $I \in \mathcal{I}$, so

$$\begin{aligned} &= L^* \int_{\mathcal{I}} \nu(I) \\ &= L^* \end{aligned}$$

which proves the claim. □

B.5. Proof of Proposition 4

Proof. By definition, the rule

$$g(x) = \arg\max_k \{\eta^k(x)\}$$

(where ties are broken in favour of smaller k) achieves the Bayes risk. In the case where all the $\eta^k(x)$ are equal there is nothing to prove, since all choices have the same probability of error. Therefore, suppose there is at least one k such that $\eta^k(x) < \eta^{g(x)}(x)$ and define

$$\begin{aligned} m(x) &= \eta^{g(x)}(x) - \max_k \{\eta^k(x) \mid \eta^k(x) < \eta^{g(x)}(x)\} \\ m_t(x) &= \eta_t^{g(x)}(x) - \max_k \{\eta_t^k(x) \mid \eta^k(x) < \eta^{g(x)}(x)\} \end{aligned}$$

The function $m(x) \geq 0$ is the margin function which measures how much better the best choice is than the second best choice, ignoring possible ties for best. The function $m_t(x)$ measures the margin of $g_t(x)$. If $m_t(x) > 0$ then $g_t(x)$ has the same probability of error as the Bayes classifier.

The assumption above guarantees that there is some ϵ such that $m(x) > \epsilon$. Using C to denote the number of classes, by making t large we can satisfy

$$\mathbb{P}\left(|\eta_t^k(X) - \eta^k(X)| < \epsilon/2\right) \geq 1 - \delta/C$$

since η_t^k is consistent. Thus

$$\mathbb{P}\left(\bigcap_{k=1}^{C} |\eta_t^k(X) - \eta^k(X)| < \epsilon/2\right) \geq 1 - K + \sum_{k=1}^{C} \mathbb{P}\left(|\eta_t^k(X) - \eta^k(X)| < \epsilon/2\right) \geq 1 - \delta$$

So with probability at least $1 - \delta$ we have

$$\begin{aligned} m_t(X) &= \eta_t^{g(X)} - \max_k \{\eta_t^k(X) \mid \eta^k(X) < \eta^{g(X)}(X)\} \\ &\geq (\eta^{g(X)} - \epsilon/2) - \max_k \{\eta_t^k(X) + \epsilon/2 \mid \eta^k(X) < \eta^{g(x)}(X)\} \\ &= \eta^{g(X)} - \max_k \{\eta^k(X) \mid \eta^k(X) < \eta^{g(X)}(X)\} - \epsilon \\ &= m(X) - \epsilon \\ &> 0 \end{aligned}$$

Since δ is arbitrary this means that the risk of g_t converges in probability to the Bayes risk. □

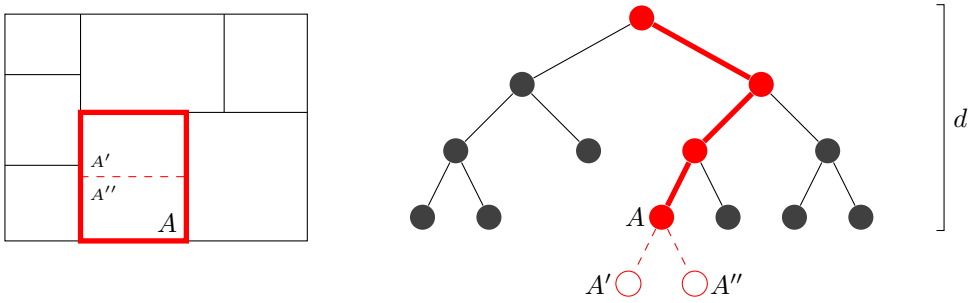

Figure 6. This Figure shows the setting of Proposition 8. Conditioned on a partially built tree we select an arbitrary leaf at depth d and an arbitrary candidate split in that leaf. The proposition shows that, assuming no other split for A is selected, we can guarantee that the chosen candidate split will occur in bounded time with arbitrarily high probability.

B.6. Proof of Theorem 1

The proof of Theorem 1 is built in several pieces.

Proposition 8. *Fix a partitioning sequence. Let t_0 be a time at which a split occurs in a tree built using this sequence, and let g_{t_0} denote the tree after this split has been made. If A is one of the newly created cells in g_{t_0} then we can guarantee that the cell A is split before time $t > t_0$ with probability at least $1 - \delta$ by making t sufficiently large.*

Proof. Let d denote the depth of A in the tree g_{t_0} and note that $\mu(A) > 0$ with probability 1 since X has a density. This situation is illustrated in Figure 6. By construction, if the following conditions hold:

1. For some candidate split in A, the number of estimation points in both children is at least $\alpha(d)$,
2. The number of estimation points in A is at least $\beta(d)$,

then the algorithm must split A when the next structure point arrives. Thus in order to force a split we need the following sequence of events to occur:

1. A structure point must arrive in A to create a candidate split point.
2. The above two conditions must be satisfied.
3. Another structure point must arrive in A to force a split.

It is possible for a split to be made before these events occur, but assuming a split is not triggered by some other mechanism we can guarantee that this sequence of events will occur in bounded time with high probability.

Suppose a split is not triggered by a different mechanism. Define E_0 to be an event that occurs at t_0 with probability 1, and let $E_1 \leq E_2 \leq E_3$ be the times at which the above numbered events occur. Each of these events requires the previous one to have occurred and moreover, the sequence has a Markov structure, so for $t_0 \leq t_1 \leq t_2 \leq t_3 = t$ we have

$$\begin{aligned}
\mathbb{P}\left(E_1 \leq t \cap E_2 \leq t \cap E_3 \leq t \,|\, E_0 = t_0\right) &\geq \mathbb{P}\left(E_1 \leq t_1 \cap E_2 \leq t_2 \cap E_3 \leq t_3 \,|\, E_0 = t_0\right) \\
&= \mathbb{P}\left(E_1 \leq t_1 \,|\, E_0 = t_0\right) \mathbb{P}\left(E_2 \leq t_2 \,|\, E_1 \leq t_1\right) \mathbb{P}\left(E_3 \leq t_3 \,|\, E_2 \leq t_2\right) \\
&\geq \mathbb{P}\left(E_1 \leq t_1 \,|\, E_0 = t_0\right) \mathbb{P}\left(E_2 \leq t_2 \,|\, E_1 = t_1\right) \mathbb{P}\left(E_3 \leq t_3 \,|\, E_2 = t_2\right) \quad .
\end{aligned}$$

We can rewrite the first and last term in more friendly notation as

$$\begin{aligned}
\mathbb{P}\left(E_1 \leq t_1 \,|\, E_0 = t_0\right) &= \mathbb{P}\left(N^s_{t_0,t_1}(A) \geq 1\right) \quad , \\
\mathbb{P}\left(E_3 \leq t_3 \,|\, E_2 = t_2\right) &= \mathbb{P}\left(N^s_{t_2,t_3}(A) \geq 1\right) \quad .
\end{aligned}$$

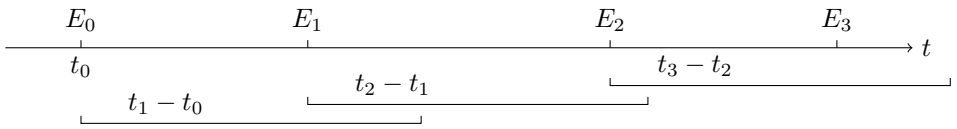

Figure 7. This Figure diagrams the structure of the argument used in Propositions 8 and 9. The indicated intervals are show regions where the next event must occur with high probability. Each of these intervals is finite, so their sum is also finite. We find an interval which contains all of the events with high probability by summing the lengths of the intervals for which we have individual bounds.

Lemma 7 allows us to lower bound both of these probabilities by $1-\epsilon$ for any $\epsilon > 0$ by making $t_1 - t_0$ and $t_3 - t_2$ large enough that

$$N^s_{t_0,t_1} \geq \frac{2}{\mu(A)} \max\left\{1,\, \mu(A)^{-1} \log\left(\frac{1}{\epsilon}\right)\right\}$$

and

$$N^s_{t_2,t_3} \geq \frac{2}{\mu(A)} \max\left\{1,\, \mu(A)^{-1} \log\left(\frac{1}{\epsilon}\right)\right\}$$

respectively. To bound the centre term, recall that $\mu(A') > 0$ and $\mu(A'') > 0$ with probability 1, and $\beta(d) \geq \alpha(d)$ so

$$\begin{aligned}\mathbb{P}\left(E_2 \leq t_2 \,|\, E_1 = t_1\right) &\geq \mathbb{P}\left(N^e_{t_1,t_2}(A') \geq \beta(d) \cap N^e_{t_1,t_2}(A'') \geq \beta(d)\right) \\ &\geq \mathbb{P}\left(N^e_{t_1,t_2}(A') \geq \beta(d)\right) + \mathbb{P}\left(N^e_{t_1,t_2}(A'') \geq \beta(d)\right) - 1 \ ,\end{aligned}$$

and we can again use Lemma 7 lower bound this by $1-\epsilon$ by making $t_2 - t_1$ sufficiently large that

$$N^e_{t_1,t_2} \geq \frac{2}{\min\{\mu(A'), \mu(A'')\}} \max\left\{\beta(d),\, \min\{\mu(A'), \mu(A'')\}^{-1} \log\left(\frac{2}{\epsilon}\right)\right\}$$

Thus by setting $\epsilon = 1 - (1-\delta)^{1/3}$ can ensure that the probability of a split before time t is at least $1-\delta$ if we make

$$t = t_0 + (t_1 - t_0) + (t_2 - t_1) + (t_3 - t_2)$$

sufficiently large. □

Proposition 9. *Fix a partitioning sequence. Each cell in a tree built based on this sequence is split infinitely often in probability. i.e all $K > 0$ and any x in the support of X,*

$$\mathbb{P}\left(A_t(x) \text{ has been split fewer than } K \text{ times}\right) \to 0$$

as $t \to \infty$.

Proof. For an arbitrary point x in the support of X, let E_k denote the time at which the cell containing x is split for the kth time, or infinity if the cell containing x is split fewer than k times (define $E_0 = 0$ with probability 1). Now define the following sequence:

$$\begin{aligned}t_0 &= 0 \\ t_i &= \min\{t \,|\, \mathbb{P}\left(E_i \leq t \,|\, E_{i-1} = t_{i-1}\right) \geq 1 - \epsilon\}\end{aligned}$$

and set $T_\delta = t_k$. Proposition 8 guarantees that each of the above t_i's exists and is finite. Compute,

$$\begin{aligned}
\mathbb{P}(E_k \le T_\delta) &= \mathbb{P}\left(\bigcap_{i=1}^{k}[E_i \le T_\delta]\right) \\
&\ge \mathbb{P}\left(\bigcap_{i=1}^{k}[E_i \le t_i]\right) \\
&= \prod_{i=1}^{k}\mathbb{P}\left(E_i \le t_i \,|\, \bigcap_{j<i}[E_j \le t_j]\right) \\
&= \prod_{i=1}^{k}\mathbb{P}(E_i \le t_i \,|\, E_{i-1} \le t_{i-1}) \\
&\ge \prod_{i=1}^{k}\mathbb{P}(E_i \le t_i \,|\, E_{i-1} = t_{i-1}) \\
&\ge (1-\epsilon)^k
\end{aligned}$$

where the last line follows from the choice of t_i's. Thus for any $\delta > 0$ we can choose T_δ to guarantee $\mathbb{P}(E_k \le T_\delta) \ge 1-\delta$ by setting $\epsilon = 1-(1-\delta)^{1/k}$ and applying the above process. We can make this guarantee for any k which allows us to conclude that $\mathbb{P}(E_k \le t) \to 1$ as $t \to \infty$ for all k as required. □

Proposition 10. *Fix a partitioning sequence. Let $A_t(X)$ denote the cell of g_t (built based on the partitioning sequence) containing the point X. Then* $\operatorname{diam}(A_t(X)) \to 0$ *in probability as $t \to \infty$.*

Proof. Let $V_t(x)$ be the size of the first dimension of $A_t(x)$. It suffices to show that $\mathbb{E}[V_t(x)] \to 0$ for all x in the support of X.

Let $X_1, \ldots, X_{m'} \sim \mu|_{A_t(x)}$ for some $1 \le m' \le m$ denote the samples from the structure stream that are used to determine the candidate splits in the cell $A_t(x)$. Use π_d to denote a projection onto the dth coordinate, and without loss of generality, assume that $V_t = 1$ and $\pi_1 X_i \sim \text{Uniform}[0,1]$. Conditioned on the event that the first dimension is cut, the largest possible size of the first dimension of a child cell is bounded by

$$V^* = \max(\max_{i=1}^{m}\pi_1 X_i, 1 - \min_{i=1}^{m}\pi_1 X_i) \ .$$

Recall that we choose the number of candidate dimensions as $\min(1+\text{Poisson}(\lambda), D)$ and select that number of distinct dimensions uniformly at random to be candidates. Define the following events:

$$\begin{aligned}
E_1 &= \{\text{There is exactly one candidate dimension}\} \\
E_2 &= \{\text{The first dimension is a candidate}\}
\end{aligned}$$

Then using V' to denote the size of the first dimension of the child cell,

$$\begin{aligned}
\mathbb{E}[V'] &\le \mathbb{E}[\mathbb{I}\{(E_1 \cap E_2)^c\} + \mathbb{I}\{E_1 \cap E_2\}V^*] \\
&= \mathbb{P}(E_1^c) + \mathbb{P}(E_2^c|E_1)\mathbb{P}(E_1) + \mathbb{P}(E_2|E_1)\mathbb{P}(E_1)\mathbb{E}[V^*] \\
&= (1-e^{-\lambda}) + (1-\frac{1}{d})e^{-\lambda} + \frac{1}{d}e^{-\lambda}\mathbb{E}[V^*] \\
&= 1 - \frac{e^{-\lambda}}{D} + \frac{e^{-\lambda}}{D}\mathbb{E}[V^*] \\
&= 1 - \frac{e^{-\lambda}}{D} + \frac{e^{-\lambda}}{D}\mathbb{E}\left[\max(\max_{i=1}^{m}\pi_1 X_i, 1 - \min_{i=1}^{m}\pi_1 X_i)\right] \\
&= 1 - \frac{e^{-\lambda}}{D} + \frac{e^{-\lambda}}{D}\cdot\frac{2m+1}{2m+2} \\
&= 1 - \frac{e^{-\lambda}}{2D(m+1)}
\end{aligned}$$

Iterating this argument we have that after K splits the expected size of the first dimension of the cell containing x is upper bounded by

$$\left(1 - \frac{e^{-\lambda}}{2D(m+1)}\right)^K$$

so it suffices to have $K \to \infty$ in probability, which we know to be the case from Proposition 9. □

Proposition 11. *Fix a partitioning sequence. In any tree built based on this sequence, $N^e(A_t(X)) \to \infty$ in probability.*

Proof. It suffices to show that $N^e(A_t(x)) \to \infty$ for all x in the support of X. Fix such an x, by Proposition 9 we can make the probability $A_t(x)$ is split fewer than K times arbitrarily small for any K. Moreover, by construction immediately after the K-th split is made the number of estimation points contributing to the prediction at x is at least $\alpha(K)$, and this number can only increase. Thus for all K we have that $\mathbb{P}\left(N^e(A_t(x)) < \alpha(K)\right) \to 0$ as $t \to \infty$ as required. □

We are now ready to prove our main result. All the work has been done, it is simply a matter of assembling the pieces.

Proof (of Theorem 1). Fix a partitioning sequence. Conditioned on this sequence the consistency of each of the class posteriors follows from Theorem 5. The two required conditions where shown to hold in Propositions 10 and 11. Consistency of the multiclass tree classifier then follows by applying Proposition 4.

To remove the conditioning on the partitioning sequence, note that Lemma 6 shows that our tree generation mechanism produces a partitioning sequence with probability 1. Apply Proposition 3 to get unconditional consistency of the multiclass tree.

Proposition 2 lifts consistency of the trees to consistency of the forest, establishing the desired result. □

B.7. Extension to a Fixed Size Fringe

Proving consistency is preserved with a fixed size fringe requires more precise control over the relationship between the number of estimation points seen in an interval, $N^e_{t_0,t}$, and the total number of splits which have occurred in the tree, K. The following two lemmas provide the control we need.

Lemma 12. *Fix a partitioning sequence. If K is the number of splits which have occurred at or before time t then for all $M > 0$*

$$\mathbb{P}\left(K \leq M\right) \to 0$$

in probability as $t \to \infty$.

Proof. Denote the fringe at time t with F_t which has max size $|F|$, and the set of leafs at time t as L_t with size $|L_t|$. If $|L_t| < |F|$ then there is no change from the unbounded fringe case, so we assume that $|L_t| \geq |F|$ so that for all t there are exactly $|F|$ leafs in the fringe.

Suppose a leaf $A_1 \in F_{t_0}$ for some t_0 then for every $\delta > 0$ there is a finite time t_1 such that for all $t \geq t_1$

$$\mathbb{P}\left(A_1 \text{ has not been split before time } t\right) \leq \frac{\delta}{|F|}$$

Now fix a time t_0 and $\delta > 0$. For each leaf $A_i \in F_{t_0}$ we can choose t_i to satisfy the above bound. Set $t = \max_i t_i$ then the union bound gives

$$\mathbb{P}\left(K \leq |F| \text{ at time } t\right) \leq \delta$$

Iterate this argument $\lceil M/|F| \rceil$ times with $\delta = \epsilon / \lceil M/|F| \rceil$ and apply the union bound again to get that for sufficiently large t

$$\mathbb{P}(K \leq M) \leq \epsilon$$

for any $\epsilon > 0$. □

Lemma 13. *Fix a partitioning sequence. If K is the number of splits which have occurred at or before time t then for any $t_0 > 0$, $K/N^e_{t_0,t} \to 0$ as $t \to \infty$.*

Proof. First note that $N^e_{t_0,t} = N^e_{0,t} - N^e_{0,t_0-1}$ so

$$\frac{K}{N^e_{t_0,t}} = \frac{K}{N^e_{0,t} - N^e_{0,t_0-1}}$$

and since N^e_{0,t_0-1} is fixed it is sufficient to show that $K/N^e_{0,t} \to 0$. In the following we write $N = N^e_{0,t}$ to lighten the notation.

Define the cost of a tree T as the minimum value of N required to construct a tree with the same shape as T. The cost of the tree is governed by the function $\alpha(d)$ which gives the cost of splitting a leaf at level d. The cost of a tree is found by summing the cost of each split required to build the tree.

Note that no tree on K splits is cheaper than a tree of max depth $d = \lceil \log_2(K) \rceil$ with all levels full (except possibly the last, which may be partially full). This is simple to see, since $\alpha(d)$ is an increasing function of d, meaning it is never more expensive to add a node at a lower level than a higher one. Thus we assume wlog that the tree is full except possibly in the last level.

When filling level d of the tree, each split incurs a cost of at least $2\alpha(d+1)$ points. This also tells us that filling level d requires that N increase by at least $2^d\alpha(d)$ (filling level d corresponds to splitting each of the 2^{d-1} leafs on level $d-1$). Filling the first d levels incurs a cost of at least

$$N_d = \sum_{k=1}^{d} 2^k \alpha(k)$$

points. When $N = N_d$ the tree can be at most a full binary tree of depth d, meaning that $K \leq 2^d - 1$.

The above argument gives a collection of linear upper bounds on K in terms of N. We know that the maximum growth rate is linear between $(N_d, 2^d - 1)$ and $(N_{d+1}, 2^{d+1} - 1)$ so for all d we can find that since

$$\frac{(2^{d+1}-1) - (2^d - 1)}{(N_{d+1}) - (N_d)} = \frac{2^{d+1} - 2^d}{\sum_{k=1}^{d+1} 2^k\alpha(k) - \sum_{k=1}^{d} 2^k\alpha(k)} = \frac{2^d}{2^{d+1}\alpha(d+1)} = \frac{1}{2\alpha(d+1)}$$

we have that for all N and d,

$$K \leq \frac{1}{2\alpha(d+1)} N + C(d)$$

where $C(d)$ is given by

$$C(d) = 2^d - 1 - \frac{1}{2}\sum_{k=1}^{d} 2^k \frac{\alpha(k)}{\alpha(d+1)} \quad .$$

From this we have

$$\frac{K}{N} \leq \frac{1}{2\alpha(d+1)} + \frac{1}{N}\left(2^d - 1 - \frac{1}{2}\sum_{k=1}^{d} 2^k \frac{\alpha(k)}{\alpha(d+1)}\right) \quad ,$$

so if we choose d to make $1/\alpha(d+1) \leq \delta/2$ and then pick N such that $C(d)/N \leq \delta/2$ we have $K/N \leq \delta$ for arbitrary $\delta > 0$ which proves the claim. □

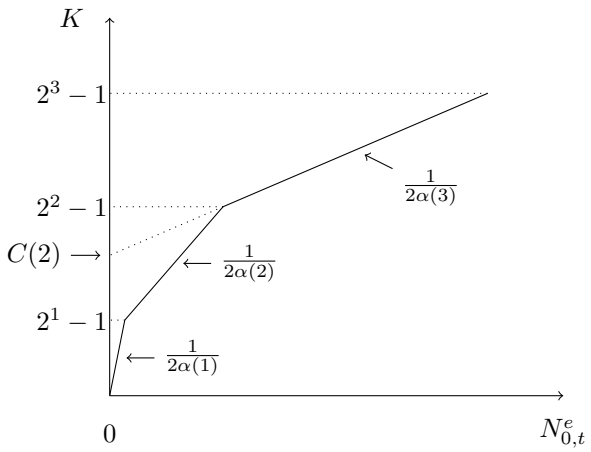

Figure 8. Diagram of the bound in Lemma 13. The horizontal axis is the number of estimation points seen at time t and the vertical axis is the number of splits. The first bend is the earliest point at which the root of the tree could be split, which requires $2\alpha(1)$ points to create 2 new leafs at level 1. Similarly, the second bend is the point at which all leafs at level 1 have been split, each of which requires at least $2\alpha(2)$ points to create a pair of leafs at level 2.

In order to show that our algorithm remains consistent with a fixed size fringe we must ensure that Proposition 8 does not fail in this setting. Interpreted in the context of a finite fringe, Proposition 8 says that any cell in the fringe will be split in finite time. This means that to ensure consistency we need only show that any inactive point will be added to the fringe in finite time.

Remark 14. *If $s(A) = 0$ for any leaf then we know that $e(A) = 0$, since $\mu(A) > 0$ by construction. If $e(A) = 0$ then $\mathbb{P}(g(X) \neq Y \mid X \in A) = 0$ which means that any subdivision of A has the same asymptotic probability of error as leaving A in tact. Our rule never splits A and thus fails to satisfy the shrinking leaf condition, but our predictions are asymptotically the same as if we had divided A into arbitrarily many pieces so this doesn't matter.*

Proposition 15. *Every leaf with $s(A) > 0$ will be added to the fringe in finite time with arbitrarily high probability.*

Proof. Pick an arbitrary time t_0 and condition on everything before t_0. For an arbitrary node $A \subset \mathbb{R}^D$, if A' is a child of A then we know that if $\{U_i\}_{i=1}^{Dm}$ are iid on $[0, 1]$ then

$$\begin{aligned}\mathbb{E}[\mu(A')] &\leq \mu(A)\mathbb{E}\left[\max_{i=1}^{Dm}(\max(U_i, 1 - U_i))\right] \\ &= \mu(A)\left(\frac{2Dm+1}{2Dm+2}\right)\end{aligned}$$

since there are at most D candidate dimensions and each one accumulates at most m candidate splits. So if A^K is any leaf created by K splits of A then

$$\mathbb{E}\left[\mu(A^K)\right] \leq \mu(A)\left(\frac{2Dm+1}{2Dm+2}\right)^K$$

Notice that since we have conditioned on the tree at t_0 so,

$$\mathbb{E}\left[\hat{p}(A^K)\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{p}(A^K) \mid \mu(A^K)\right]\right] = \mathbb{E}\left[\mu(A^K)\right] \quad .$$

We can bound $\hat{p}(A^K)$ with Hoeffding's inequality,

$$\mathbb{P}\left(\hat{p}(A^K) \geq \mu(A)\left(\frac{2Dm+1}{2Dm+2}\right)^K + \epsilon\right) \leq \exp\left(-2|A^K|\epsilon^2\right) \quad .$$

Set $(2^{K+1}|L|)^{-1}\delta = \exp\left(-2|A^K|\epsilon^2\right)$ and invert the bound to get

$$\mathbb{P}\left(\hat{p}(A^K) \geq \mu(A)\left(\frac{2Dm+1}{2Dm+2}\right)^K + \sqrt{\frac{1}{2|A^K|}\log\left(\frac{2^{K+1}|L|}{\delta}\right)}\right) \leq \frac{\delta}{2^{K+1}|L|}$$

Pick an arbitrary leaf A_0 which is in the tree at time t_0. We can use the same approach to find a lower bound on $\hat{s}(A_0)$:

$$\mathbb{P}\left(\hat{s}(A_0) \leq s(A_0) - \sqrt{\frac{1}{2|A_0|}\log\left(\frac{2^{K+1}|L|}{\delta}\right)}\right) \leq \frac{\delta}{2^{K+1}|L|}$$

To ensure that $\hat{s}(A_0) \geq \hat{p}(A^K)$ $(\geq \hat{s}(A^K))$ fails to hold with probability at most $\delta 2^{-K}|L|^{-1}$ we must choose K and t to make

$$s(A_0) \geq \mu(A)\left(\frac{2Dm+1}{2Dm+2}\right)^K + \sqrt{\frac{1}{2|A^K|}\log\left(\frac{2^{K+1}|L|}{\delta}\right)} + \sqrt{\frac{1}{2|A_0|}\log\left(\frac{2^{K+1}|L|}{\delta}\right)}$$

The first term goes to 0 as $K \to \infty$. We know that $|A^K| \geq \alpha(K)$ so the second term also goes to 0 provided that $K/\alpha(K) \to 0$, which we require.

The third term goes to 0 if $K/|A_0| \to 0$. Recall that $|A_0| = N^e_{t_0,t}(A_0)$ and for any $\gamma > 0$

$$\mathbb{P}\left(N^e_{t_0,t}(A) \leq N^e_{t_0,t}\mu(A) - \sqrt{\frac{1}{2N^e_{t_0,t}}\log\left(\frac{1}{\gamma}\right)}\right) \leq \gamma$$

From this we see it is sufficient to have $K/N^e_{t_0,t} \to 0$ which we established in a lemma.

In summary, there are $|L|$ leafs in the tree at time t_0 and each of them generates at most 2^K different A^K's. Union bounding over all these leafs and over the probability of $N^e_{t_0,t}(A_0)$ growing sublinearly in $N^e_{t_0,t}$ we have that, conditioned on the event that A_0 has not yet been split, A_0 is the leaf with the highest value of $\hat{s}$ with probability at least $1 - \delta - \gamma$ in finite time. Since δ and γ are arbitrary we are done. □