1 Introduction

Density estimation is one of the fundamental tasks in statistics and machine learning, as it often appears as a subroutine in tasks such as regression, classification, and clustering. However, density estimation in high dimensions remains a challenging problem due to the curse of dimensionality, which affects the convergence rates of many popular density estimation techniques. In addition, common density estimation techniques often lack in interpretability, adaptability, speed, accuracy, and reasonableness of assumptions. For instance, mixtures of Gaussians (MoGs) require parametric assumptions, fixed bandwidth kernel density estimators (KDEs) can’t adapt well to heterogeneous smoothness (even within even a single dimension), and adaptive bandwidth kernel density estimators (which seek to be more adaptable) are complex and expensive computationally.

A simple but seemingly novel idea for density estimation is that of a piecewise-constant density estimator, which can be imagined as a histogram with adaptive bandwidths for each dimension. In a 2011 paper entitled ‘Density Estimation Trees’, Ram and Gray introduce a technique called the Density Estimation Tree (DET) which creates such an estimator [10]. The motivation comes from the widespread and successful use of classification and regression trees in supervised learning; while trees are not the most accurate estimators for classification or regression, they are heavily used because they are intuitive, interpretable, and adapt to heterogeneity within and between dimensions. Additional benefits of trees include automatic feature selection and efficient test querying. DETs bring the idea of classification and regression trees to density estimation with the goal of increased interpretability and adaptability.

While work prior to Ram and Gray used similar hierarchical classification or regression ideas in density estimation [5, 13, 4, 9, 11], none used trees directly as density estimators, and in our understanding, Ram and Gray is the first paper to explicitly outline the construction of such a piecewise-constant adaptive density estimator. The construction of DETs is based upon the classical tree learning algorithm by Breiman with modifications for the unsupervised density estimation setting [2]. Ram and Gray prove consistency of DETs under mild assumptions on the density class for continuous features, but DETs can also handle categorical, ordinal, or mixed data, though consistency has not been established in these cases. Ram and Gray demonstrate through experiments that DETs adapt to local smoothness within one dimension, ignore irrelevant dimensions, and are highly interpretable (even if not as accurate as the more computationally expensive adaptive KDE methods).

Two other recent papers propose piecewise-constant density estimators which are qualitatively very similar to DETs, though the exact details of construction differ. In a 2014 arXiv preprint paper entitled ‘Density Estimation via Adaptive Partition and Discrepancy Control’, Yang and Wong discuss an estimator defined by building up a binary partition of a hyper rectangle $\in \mathbb{R}^d$ sequentially, using discrepancy criteria at each step to determine the appropriate stopping point [15]. The authors prove their estimator converges uniformly at order $O(1/\sqrt{N})$. Interestingly, Yang and Wong compare their estimator to two recent Bayesian methods which they state are roughly analogous [14, 8]. We
do not explore those Bayesian methods here, though it appears that a thorough comparison may not have been performed and provides an interesting challenge.

Finally, in a 2014 arXiv preprint paper entitled ‘Multivariate Density Estimation Based on Adaptive Partitioning: Convergence Rate, Variable Selection and Spatial Adaptation’, Liu and Wong seek to characterize the set of all piecewise constant estimators that are built up through binary partitions of the data [7]. This work provides a more general characterization of function spaces which can be approximated by such piecewise-constant estimators and uses some interesting but challenging theory to establish results. The main finding of this paper is that the convergence rate of this type of estimator is dimension-free, implying that under certain assumptions and constraints, piecewise-constant estimators can avoid the curse of dimensionality which plagued other estimators (such as histograms and KDEs) in high dimensions.

In the remainder of this survey, we will discuss the main theoretical contributions of each of these three key papers. We will first treat each paper separately by outlining the notation and assumptions, key results, and giving brief proof outlines for selected results in each work. Comparisons of the different approaches and key results are given in the conclusion. In comparing the approaches of three different groups of authors, we hope to give a unified picture of the current theory surrounding piecewise-constant adaptive density estimation techniques.

2 Ram and Gray (2011)

2.1 Notation and Assumptions

In the problem of density estimation, we want to estimate the unknown density function \( f \) on \( \mathcal{X} \) as closely as possible with \( \hat{f} \) : \( \mathcal{X} \rightarrow \mathbb{R}_+ \). To estimate this function, we will use \( n \) i.i.d. observations \((X_1, X_2, ..., X_n) \in \mathcal{X} \) drawn from the unknown density \( f \).

Ram and Gray use the following definition as their Density Estimation tree:

**Definition 1.** The piecewise constant density estimate of the decision tree \( T \) built with \( N \in \mathbb{R}^d \) observations is:

\[
\hat{f}_N(x) = \sum_{l \in \hat{T}} \frac{|l|}{NV_l} I(x \in l)
\]

Where \( \hat{T} \) is the set of leaves, \(|l|\) is the number of observations in leaf \( l \), and \( V_l \) is the volume of leaf \( l \) with a \( d \) dimensional bounding box.

They used the common Integrated Squared Error loss function to evaluate their estimator:

**Definition 2.** Integrated Squared Error (ISE) loss function measuring the distance between the estimated and true density:

\[
\int_{\mathcal{X}} (\hat{f}_N(x) - f(x))^2 dx
\]

2.2 Building the Tree

The first step the authors made to create the tree was to rewrite the loss function into a more convenient form. Using standard techniques such as expanding the square, using a plug in estimator and Monte Carlo substitution, (ie: \( \int_{\mathcal{X}} \hat{f}_N(x)f(x)dx = \frac{1}{N} \sum_{i=1}^{N} \hat{f}_N(X_i)(X_i) \)), they were able to rewrite the estimator for of the ISE for the Density tree from the minimization problem

\[
\min_{f_n \in \mathbb{R}_+}[\int_{\mathcal{X}} (\hat{f}_N(x))^2 dx - \frac{2}{N} \sum_{i=1}^{N} \hat{f}_N(X_i)]
\]

as:

\[
R(T) = \sum_{l \in \hat{T}} \left[ \frac{|l|^2}{N^2V_l} \right]
\]
From here, they were able to define the greedy surrogate of the error for any node is:

$$R(t) = -\frac{|t|^2}{N^2 V_t}$$

The tree is then built using the same tree learning algorithm in from Breiman [2], maximizing the reduction in this greedy surrogate error based on the $N$ observations.

Formally, the algorithm is:

1. **Splitting.** First, we split each node into two children, calling $S$ the family off all univariate splits, and $s$ the split at each node $t$. Then the best split is, as we’ve said above, the one which maximally reduces the loss function: $\sup_{s \in S} [R(t) - R(t_l) - R(t_r)]$

2. **Pruning.** To avoid overfitting, we prune the tree by introducing a regularization parameter to penalize the complexity of the tree. Thus, the error of a tree rooted at node $t$ is $R_\alpha(t) = R(\hat{t}) + \alpha |\hat{t}|$ where $\hat{t}$ is the set of leaves rooted at node $t$.

3. **Cross Validation.** Finally, we choose the optimal $\alpha$ using leave one out cross validation (LOO-CV). This is given by:

$$\hat{J}(\alpha) = \int_X \hat{f}_N^\alpha(x)^2 dx - \frac{2}{N} \sum_{i=1}^N \hat{f}_i^\alpha(X_i)$$

Where the first term on the RHS is the decision tree pruned with regularization parameter $\alpha$ and the second $\hat{f}_N^\alpha$ is the same, except without training example $X_i$. Thus, the optimal tree is the alpha that minimizes the above equation: $\alpha^* = \inf_{\alpha} \hat{J}(\alpha)$

Thus, after completing these three steps, we have our estimate of the tree $T$ and its optimal size.

### 2.3 Key Results

The main theoretical result from Ram and Gray is that the DET is consistent under mild assumptions on the model class of the unknown distribution $f$. They prove the consistency for continuous features, as opposed to categorical/ordinal ones. However, there is a section in their paper that describes what the density would look like if we were using categorical/ordinal features.

Formally, the theorem is:

**Theorem 1.** The estimator $\hat{f}_N$ defined in definition 1 satisfies:

$$\Pr\left( \lim_{N \to \infty} \int_X (\hat{f}_N(x) - f(x))^2 dx = 0 \right) = 1$$

Which says that the probability that the loss function will go to 0 as $N \to \infty$ is 1.

### 2.4 Proof Outlines

**Proof.** [Proof of Theorem] Proving the Consistency of the DET with continuous features is similar to the proof regarding the consistency of of regression trees in [2]. Using using the ISE loss function from definition 2, we need to show:

$$\Pr\left( \lim_{N \to \infty} \int_X (\hat{f}_N(x) - f(x))^2 dx = 0 \right) = 1$$

To start, they defined the model class in which the estimator holds. Let $B = \{ t \in X : b' x \leq c, b \in \mathbb{R}^d, c \in \mathbb{R} \}$ that are solutions to a system of $d$ inequalities, where $d$ is a positive integer. Then, we
have that for every leaf \( l \in \hat{T} \), we have a solution to the above inequalities, putting one entry equal to 1 and the rest equal to zero. The result of this, gives us \( \hat{T} \in B \).

Next, Ram and Gray introduce the empirical distribution function (which looks similar to the DET), and used the well known Glivenko Cantelli theorem to introduce a uniform convergence equation. From here, they relate the empirical distribution function to the DET estimator, using the fact that

\[
\hat{F}_N(t) = \int_t f_N(x) dx
\]

To arrive at the following equation:

\[
= \mathbb{P} \left( \lim_{N \to \infty} \sup_{t \in B} \int_t \hat{f}_N(x) - f(x) dx = 0 \right) = 1
\]

\[
\Rightarrow \mathbb{P} \left( \lim_{N \to \infty} \sup_{t \in B} \int_t \hat{f}_N(x) - f(x) dx \geq 0 \right) = 1
\]

The final ingredient to the proof is the assumption that as you take the number of data-points to infinity, the diameter size of each unit in the estimator gets arbitrarily small, ie:

\[
\lim_{N \to \infty} \mathbb{P}(\text{diameter}(t) \geq \epsilon) = 0.
\]

This gives us that \( \mathbb{P}(\lim_{N \to \infty} \int_t dx = 0) = 1 \), which we can use to say that for \( x' \in t \):

\[
\lim_{N \to \infty} \sup_{t \in B} \int_t |\hat{f}_N(x) - f(x)| dx \leq \lim_{N \to \infty} \int_t |\hat{f}_N(x') - f(x')| dx' = 0
\]

So we have introduced a model class, related the DET to the empirical distribution function and used the Glivenko Cantelli theorem in order to get an uniform bound equation which looks similar to our loss function, and assumed that the size of the leaf nodes gets smaller and smaller as the datapoints grow to arrive at the conclusion:

\[
\mathbb{P} \left( \lim_{N \to \infty} \sup_{t \in B} \int_t (\hat{f}_N(x) - f(x))^2 dx = 0 \right) = 1
\]

which was the desired result.

\( \square \)

3 Yang and Wong (2014)

3.1 Notation and Assumptions

Discrepancy is a set of criteria to measure the uniformity of points in the unit hypercube \([0, 1]^d\), which is used to determine when to create a new split in the partition underlying the estimator.

**Definition 3.** (Classic star discrepancy)

\[
D_n^*(x_1, \ldots, x_n) = \sup_{a \in [0, 1]^d} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{x_i \in [0, a]} - \prod_{i=1}^d a_i \right|
\]

Total variation is a way to measure the variability of a function on a particular interval which is used to characterize the class of densities for which the estimator will achieve the consistency results.

**Definition 4.** (Hardy-Krause total variation measure)

For any hyper-rectangle \([a, b]\) such that all of the involved partial derivatives of \( f \) are continuous on \([a, b] \),

\[
V_{HK}^{[a,b]}(f) = \sum_{u \subseteq \{1, \ldots, d\}} \left\| \frac{\partial^{\| u \|} f}{\partial x_u} \bigg|_{x_{-u} = b_{-u}} \right\|_1
\]
To define the estimator and prove its properties, it is assumed only that the density function \( p \) is defined on a hyper-rectangle \([0, 1]^d\) with \( V^{[0,1]^d}_{HK}(f) < \infty\).

**Definition 5.** Let \( x_1, \ldots, x_N \in [0, 1]^d \) be an iid sample drawn from the distribution \( p(x) \) defined on \([0, 1]^d\). For a split of \([0, 1]^d\) into sub rectangles \( \{[a_i, b_i]\}_{i=1}^N \), define a piecewise constant density estimator

\[
\hat{p}(x) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^d n_i \prod_{j=1}^d (b_{ij} - a_{ij}) I(x \in [a_i, b_i])
\]

### 3.2 Key Results

As defined above, a piecewise-constant density estimator is built of sequential binary partitions of a hyper-rectangle in \( \mathbb{R}^d \). The motivation for using the star discrepancy as a way to decide on further splitting of the space (and the assumption on the total variation) comes from the following well-known result:

**Theorem 2.** (Koksma-Hlawka Inequality) Let \( x_1, \ldots, x_n \in [0, 1]^d \) and \( f \) be defined on \([0, 1]^d\). Then

\[
\left| \int_{[0,1]^d} f(x)dx - \frac{1}{n} \sum_{i=1}^n f(x_i) \right| \leq D_n^*(x_1, \ldots, x_n) V^{[0,1]^d}_{HK}(f)
\]

The goal is to decide when to further split a sub-rectangle \([a, b] = \prod_{j=1}^d [a_j, b_j]\) based on a threshold of the discrepancy function, so the authors translate Theorem 2 into this specific context:

**Theorem 3.** Let \( P = \{x_1, \ldots, x_n \in [a, b]\} \) and \( f \) be defined on the \( d \)-dimensional hyper-rectangle \([a, b]\). Let \( \hat{P} = \{\tilde{x}_i = (\frac{x_{i1} - a_1}{b_1}, \ldots, \frac{x_{id} - a_d}{b_d})\}_{i=1}^n \). Then

\[
\left| \int_{[a,b]} f(x)dx - \prod_{i=1}^d (b_i - a_i) \frac{n}{n} \sum_{i=1}^n f(x_i) \right| \leq \prod_{i=1}^n (b_i - a_i) D_n^*(\hat{P}) V^{[a,b]}_{HK}(f)
\]

Finally, we state without proof two lemmas used to establish the key result.

**Theorem 4.** Let \( f \) be defined on a hyper-rectangle \([a, b]\). Let \( \{[a_i, b_i] : 1 \leq i \leq m < \infty\} \) be a split of \([a, b]\). Then

\[
\sum_{i=1}^m V^{[a_i, b_i]}_{HK}(f) = V^{[a,b]}_{HK}(f)
\]

**Theorem 5.** Let \( D_{n,d}^* = \inf_{x_1, \ldots, x_n \in [0,1]^d} D_n^*(x_1, \ldots, x_n) \). Then \( D_n^* \leq cd^{1/2}n^{-1/2} \).

The key theorem gives a consistency result.

**Theorem 6.** Consider the piecewise constant estimator \( \hat{p} \) in Definition 5

Let \( P = \{x_1, \ldots, x_N \in [a_i, b_i] \in N^+\} \) be the points in each sub-region.

Let \( \hat{P} = \{\tilde{x}_i = (\frac{x_{i1} - a_1}{b_1}, \ldots, \frac{x_{id} - a_d}{b_d})\}_{i=1}^N \). In each sub-region \([a_i, b_i]\), \( P \) satisfies

\[
D_{n_i,d}^*(\tilde{P}) \leq \theta \sqrt{\frac{N}{n_i d}} D_{n_i,d}^*
\]

for a positive constant \( \theta \). Then

\[
\left| \int_{[0,1]^d} f(x) \hat{p}(x)dx - \frac{1}{N} \sum_{i=1}^N f(x_i) \right| \leq \frac{\theta}{\sqrt{N}} V^{[0,1]^d}_{HK}(f)
\]

Following from this result a uniform convergence rate for the distribution can be derived.

**Theorem 7.** For any hyper-rectangle \( A = [a, b] \subset [0, 1]^d \), let \( \hat{P}(A) = \int_A \hat{p}(x)dx \) and \( P(A) = \int_A p(x)dx \). Then \(|P(A) - \hat{P}(A)|\) converges uniformly to 0 at order \( O(1/\sqrt{N}) \).
3.3 Proof Outlines

Here we give the outline of the proof of Theorem 6.

**Proof.** [Proof of Theorem 6] Let \( d_i = \frac{n_i}{\prod_{j=1}^{d} (b_{ij} - a_{ij})} \). Applying Theorem 3 to each \([a_i, b_i]\), we have

\[
\left| \int_{[a_i, b_i]} f(x) dx - \frac{1}{N} \sum_{i=1}^{N} f(x_i) \right| \leq \frac{1}{N} \sum_{i=1}^{d} \prod_{j=1}^{d} (b_{ij} - a_{ij}) D_{n_i}^{*} (\hat{P}_i) V_{HK}^{[a_i, b_i]} (f)
\]

By the triangle inequality and Theorems 5 and 4,

\[
\left| \int_{[0,1]^d} f(x) \hat{p}(x) dx - \frac{1}{N} \sum_{i=1}^{N} f(x_i) \right| \leq \frac{1}{N} \sum_{i=1}^{d} \prod_{j=1}^{d} (b_{ij} - a_{ij}) D_{n_i}^{*} (\hat{P}_i) V_{HK}^{[a_i, b_i]} (f)
\]

\[
\leq \frac{1}{N} \sum_{i=1}^{d} \frac{\sqrt{N} \theta}{\sqrt{n_i d c}} V_{HK}^{[a_i, b_i]} (f)
\]

\[
= \frac{\theta}{\sqrt{N}} \sum_{i=1}^{d} V_{HK}^{[a_i, b_i]} (f) = \frac{\theta}{\sqrt{N}} \sum_{i=1}^{d} V_{HK}^{[0,1]^d} (f)
\]

\( \square \)

4 Liu and Wong (2014)

4.1 Notation and Assumptions

It is assumed that the data lie in the unit cube \( \mathcal{Y} \) in \( \mathbb{R}^d \). It is assumed that the true density exists with respect to a \( \sigma \)-finite measure \( \mu \) (in this case the Lebesgue measure). Define \( \Theta \) as the collection of all density functions on this space: \( \Theta = \{ f \in m \mathcal{Y}^+: \int_{\mathcal{Y}} f d\mu = 1 \} \); it is assumed the true density lies in this class. Because \( \Theta \) is infinite-dimensional, consider a sequence of finite dimensional approximating spaces \( \Theta_1, \Theta_2, ..., \Theta_l, ... \) based on binary partitions. The total number of possible partitions after the first step is equal to the dimension \( d \). This leads to the following definition of the approximating space:

**Definition 6.** Let \( \bigcup_{i=1}^{l} A_i \) be a binary partition of \( A \) of size \( I \). Then define the set of density functions supported on binary partitions of size \( I \) as

\[
\Theta_I = \left\{ f \in \Theta : f = \sum_{i=1}^{l} \beta_i \mathbb{1}_{A_i}, \sum_{i=1}^{l} \beta_i \mu(A_i) = 1 \right\}
\]

Let the distance between density functions in \( \Theta \) or \( \Theta_I \) be the Hellinger distance:

**Definition 7.** (Hellinger distance)

\[
\rho(f, g) = \left( \int_{\mathcal{Y}} \left( \sqrt{f(y)} - \sqrt{g(y)} \right)^2 dy \right)^{1/2}
\]

Let \( f_1 = \sum_{i=1}^{l} \beta_1^i A_1^i \), \( f_2 = \sum_{i=1}^{l} \beta_2^i A_2^i \) be in \( \Theta_I \). The Hellinger distance is simply

\[
\rho^2(f_1, f_2) = \sum_{i=1}^{l} \sum_{j=1}^{l} \left( \sqrt{\beta_1^i} - \sqrt{\beta_2^j} \right)^2 \mu(A_1^i \cap A_2^j)
\]

Let \( \hat{\Theta} \) be the class of density functions that is well-approximated by density functions that are constant within each subregion of the binary partitions, defined as:

**Definition 8.** Suppose there exists a sequence \( \pi_I f \in \Theta \) satisfying \( \rho(\pi_I f, f) = O \left( \frac{1}{r^r} \right) \) for \( r > 0 \). Then \( f \in \hat{\Theta} \).
The key results presented by Liu and Wong are over the space $\tilde{\Theta}$, and it is claimed that this class of densities is still a rich class of functions.

**Definition 9.** (Maximum likelihood estimator for $f \in \Theta$)

$$\hat{f}_{n,I} = \arg \max_{f \in \tilde{\Theta}} \sum_{i=1}^{I} N_i \log \beta_i$$

This MLE is well-defined because given the binary partition, the underlying distribution is multinomial, allowing us to determine $\beta$ by maximizing the log likelihood over the finite number of possible partitions in $\tilde{\Theta}$. Each partition is scored by the maximum likelihood it can achieve.

To measure the complexity of densities in $\tilde{\Theta}$, the authors use metric entropy with bracketing, based on metric entropy as defined in Kolmogorov and Tikhomirov [6].

**Definition 10.** Let $(\Theta, \rho)$ be a separable pseudo-metric space. $\Theta(\epsilon)$ is a finite set of pairs of functions $\{(f_L^j, f_U^j), j = 1, \ldots, N\}$ satisfying $\rho(f_L^j, f_U^j) \leq \epsilon$, and for any $f \in \Theta$, there exists $j$ such that $f_L^j \leq f \leq f_U^j$. Let $N(\epsilon, \Theta, \rho) = \min\{\text{card}\Theta(\epsilon)\}$. Then the metric entropy with bracketing of $\Theta$ is

$$H(\epsilon, \Theta, \rho) = \log N(\epsilon, \Theta, \rho).$$

### 4.2 Key Results

In order to establish the key result, the authors first give two lemmas which bound the metric entropy of $\tilde{\Theta}$.

**Theorem 8.** Let $\Theta_1^{\cup A_i,d}$ be $f \in \Theta$ with partition $\cup_{i=1}^{I} A_i$ and $\rho(f, \pi_1 f_0) \leq d$. Then

$$H(u, \Theta_1^{\cup A_i,d}, \rho) \leq I \log I + I \log \frac{d}{u} + c$$

where $c$ is a constant that does not depend on the binary partition.

**Theorem 9.** Let $\Theta_1^d = \{f \in \Theta : \rho(f, \pi_1 f_0) \leq d\}$. Then

$$H(u, \Theta_1^d, \rho) \leq I \log p + (I + 1) \log (I + 1) + \frac{I}{2} \log I + I \log \frac{d}{u} + c$$

where $c$ is a constant that does not depend on the binary partition.

The key result is:

**Theorem 10.** Let $\hat{f}_{n,I}$ be the maximum likelihood estimator on $\Theta$ (Definition 9). When $r > 1/2$, assume that $n$ and $I$ satisfy $\frac{n}{2^r} = o(1)$ and $\frac{I \log I \log n}{n} = o(1)$. Then the convergence rate of $\hat{f}_{n,I}$ is $\max\left\{\frac{I \log I \log n}{n}, \frac{1}{r}\right\}$.

### 4.3 Proof Outlines

The proof of the key result is quite lengthy and detailed. Instead of trying to give a specific outline, we only give a high-level overview of the ideas used in the proof.

The key idea is that the optimal convergence rate can be achieved by balancing the sample size with the complexity of the approximating spaces $\Theta$. More complex $\Theta$ spaces will give more accurate estimators, but overly complex spaces will overfit data which is necessarily limited to a finite sample size. This is much the same problem as faced when trying to estimate a density function from the empirical CDF: some kind of smoothness must be imposed on the estimator class or else the MLE is simply infinite spikes at each data point.

The authors use metric entropy with bracketing as defined above as a way to measure the complexity of the spaces $\Theta$. Intuitively, this counts how many pairs of functions in an $\epsilon$-net are needed to upper and lower bound all elements of $\Theta$.

The authors also use results from studies of empirical processes, where the process is indexed by the log-likelihood ratios. As the authors state, most of the results require boundedness or finite
moment generating functions which does not always happen with log-likelihood ratios, so truncated log-likelihood ratios are used.

The key argument used to establish the convergence rate is a change-of-measure argument, so that the probability under $P_{\pi, f_0}$, which is multinomial, can be calculated explicitly.

## 5 Conclusion

We have reviewed a few different approaches to estimate densities using piecewise-constant estimators built on flexible partitions of the data. These approaches give a new perspective on how to estimate the density on a set of N i.i.d. observations, which could be immediately useful in many data analysis settings, as it’s easily implementable, and has unique and desirable properties. Similar to their success in supervised learning, the tree based density estimator provides a simple set of rules for the underlying structure of the distribution, is able to deal with different kinds of variables (continuous, categorical, etc...), and performs automatic feature selection. The other piecewise-constant partitioning approaches have been shown empirically to have similar features, though not all have been extended to non-continuous variables.

The first group to present the idea of using the tree framework for density estimation was Ram and Gray in 2011. In their paper, they defined the estimator, and showed it’s consistency when all of the observations are continuous. Along with being the first to propose the idea, this group laid out plenty of future directions for related work. For instance, they hypothesized what the convergence rate would be, along with speculating that this estimate could be a route towards avoiding the curse of dimensionality in high dimensions.

In 2014, Yang and Wong came up with a similar idea, using a piecewise constant estimator built up sequentially using binary partitions. While the structure of the resulting estimator is similar to a DET, the construction of the estimator is a bit different. The authors used discrepancy criteria to decide when to further subdivide the domain, which essentially results in the domain being divided until the data are roughly uniform on each part, then assigning a constant value on each piece of the domain which reflects the distribution of the data on that piece. Furthermore, they not only proved consistency of their piecewise constant estimator, but also shows the rate of convergence was on the order $O\left(\frac{1}{\sqrt{N}}\right)$.

Finally, Wong revisited the idea once again in 2014, this time accompanied by Liu. This paper goes after the most ambitious result, namely proving the cases where these piecewise constant estimators avoid the curse of dimensionality. The estimators considered are also built as binary partitions of the data with piecewise constant estimates using the MLE on each part of the domain. The results here are slightly more general than the previous papers in that they characterize the entire space of such binary partition estimators. With certain assumptions made about the true density (which involve restriction only to densities which are well-estimated by such piecewise-constant estimators), the convergence rate is shown to be independent of dimensions.

Density estimation is perhaps the most fundamental task of statistics/machine learning, since it’s a necessary step in classification and regression. That said, common low dimensional density estimators such as Histograms and Kernel Density estimators have issues when applied to high dimensional datasets, which are becoming more common in practice. From this, density trees are a very intriguing idea which allows sidestepping of the curse of dimensionality. The estimator is also naturally adaptable both between and within dimension, which avoids problems dealing with the bandwidth.

There are many ways to judge a density estimator, and every problem dealing with the analysis of data presents different challenges. When dealing with statistical problems, one needs to consider the pro’s and cons of different techniques, and use the one which best matches the situation. We believe the density tree defined in these three papers is potentially a very useful tool, depending on the situation the practitioner is in. For instance, if you have 20,000 characteristics on cancer and non-cancer patients, density trees could give you an idea of the variation within each of these groups, rather than just the features which best differentiate them.
A Group Member Contributions

Both group members read all three key papers discussed in this survey. In addition, both group members shared the task of performing a literature search to find other instances of similar research, and to confirm that no key pieces of work were left out. Similarly, the proposal and midway report were composed as a team. Both team members also worked on slides for the two-minute blitz talk, and we each spoke for one minute.

For the composition of the final report, we split the work as follows:

- Lee: created the body of the report, drafted introduction and conclusion, and wrote the section pertaining to the Ram and Gray paper.
- Natalie: helped edit the introduction and conclusion and wrote the sections pertaining to the Yang and Wong and Wong and Liu papers.

References