ADAPTIVE ESTIMATION OF MULTIVARIATE PIECEWISE POLYNOMIALS AND BOUNDED VARIATION FUNCTIONS BY OPTIMAL DECISION TREES

By Sabyasachi Chatterjee* and Subhajit Goswami†

University of Illinois at Urbana-Champaign and Tata Institute of Fundamental Research

117 Illini Hall
Champaign, IL 61820
E-mail: sc1706@illinois.edu

1, Homi Bhabha Road
Colaba, Mumbai 400005, India
E-mail: goswami@math.tifr.res.in

Proposed by Donoho (1997), Dyadic CART is a nonparametric regression method which computes a globally optimal dyadic decision tree and fits piecewise constant functions in two dimensions. In this article we define and study Dyadic CART and a closely related estimator, namely Optimal Regression Tree (ORT), in the context of estimating piecewise smooth functions in general dimensions in the fixed design setup. More precisely, these optimal decision tree estimators fit piecewise polynomials of any given degree. Like Dyadic CART in two dimensions, we reason that these estimators can also be computed in polynomial time in the sample size N via dynamic programming. We prove oracle inequalities for the finite sample risk of Dyadic CART and ORT which imply tight risk bounds for several function classes of interest. Firstly, they imply that the finite sample risk of ORT of *order* $r \geq 0$ is always bounded by $Ck\frac{\log N}{N}$ whenever the regression function is piecewise polynomial of degree r on some reasonably regular axis aligned rectangular partition of the domain with at most k rectangles. Beyond the univariate case, such guarantees are scarcely available in the literature for computationally efficient estimators. Secondly, our oracle inequalities uncover minimax rate optimality and adaptivity of the Dyadic CART estimator for function spaces with bounded variation. We consider two function spaces of recent interest where multivariate total variation denoising and univariate trend filtering are the state of the art methods. We show that Dyadic CART enjoys certain advantages over these estimators while still maintaining all their known guarantees.

*Supported by NSF Grant DMS-1916375

†Supported by an IDEX grant from Paris-Saclay and partially by a grant from the Infosys Foundation

Author names are sorted alphabetically.

Keywords and phrases: Optimal Decision Trees, Dyadic CART, Piecewise Polynomial Fitting, Oracle Risk Bounds, Bounded Variation Function Estimation

1. Introduction. Decision Trees are a widely used technique for nonparametric regression and classification. Decision Trees result in interpretable models and form a building block for more complicated methods such as bagging, boosting and random forests. See Loh (2014) and references therein for a detailed review. The most prominent example of decision trees is classification and regression trees (CART), proposed by Breiman et al. (1984). CART operates in two stages. In the first stage, it recursively partitions the space of predictor variables in a greedy top down fashion. Starting from the root node, a locally optimal split is determined by an appropriate optimization criterion and then the process is iterated for each of the resulting child nodes. The final partition or decision tree is reached when a stopping criterion is met for each resulting node. In the second stage, the final tree is pruned by what is called *cost complexity pruning* where the cost of a pruned tree thus obtained is proportional to the number of the leaves of the tree; see Section 9.2 in Friedman et al. (2001) for details.

A possible shortcoming of CART is that it produces locally optimal decision trees. It is natural to attempt to resolve this by computing a globally optimal decision tree. However, computing globally optimal decision tree is computationally a hard problem. It is known (see Laurent and Rivest (1976)) that computing an optimal (in a particular sense) binary tree is NP hard. A recent paper of Bertsimas and Dunn (2017) sets up an optimization problem (see in (Bertsimas and Dunn, 2017, Equation 1)) in the context of classification, which aims to minimize (among all decision trees) misclassification error of a tree plus a penalty proportional to its number of leaves. The paper formulates this problem as an instance of mixed integer optimization (MIO) and claims that modern MIO developments allow for solving reasonably sized problems. It then demonstrates extensive experiments for simulated and real data sets where the optimal tree outperforms the usual CART. These experiments seem to provide strong empirical evidence that optimal decision trees, if computed, can perform significantly better than CART. Another shortcoming of CART is that it is typically very hard to theoretically analyze the full algorithm because of the sequence of data dependent splits. Some results (related to the current paper) exist for the subtree obtained in the pruning stage, conditional on the maximal tree grown in the first stage; see Gey and Nedelec (2005) and references therein. Theoretical guarantees for the widely used Random Forests are also typically hard to obtain in spite of much recent work; see Scornet et al. (2015), Wager and Walther (2015), Ishwaran (2015) and references therein. On the other hand, theoretical analysis for optimal decision trees can be obtained since it can be seen as penalized empirical risk minimization.

One class of decision trees for which an optimal tree can be computed efficiently, in low to moderate dimensions, is the class of *dyadic decision trees*. These trees are constructed from recursive dyadic partitioning. In the case of regression on a two-dimensional grid design, the paper Donoho (1997) proposed a penalized least squares estimator called the Dyadic CART estimator. The author showed that it is possible to compute this estimator by a fast bottom up dynamic program which has linear time computational complexity $O(n \times n)$ for a $n \times n$ grid. Moreover, the author showed that Dyadic CART satisfies an oracle risk bound which in turn was used to show that it is adaptively minimax rate optimal over classes of

anisotropically smooth bivariate functions. Ideas in this paper were later used in Nowak et al. (2004) in the context of adaptively estimating piecewise Holder smooth functions. The idea of dyadic partitioning were also used in classification in papers such as Scott and Nowak (2006) and Blanchard et al. (2007) who studied penalized empirical risk minimization over dyadic decision trees of a fixed maximal depth. They also proved oracle risk bounds and showed minimax rate optimality for appropriate classes of classifiers. Minimax rates of convergence have also been obtained for various models of dyadic classification trees in Lecué (2008). In the related problem of density estimation, dyadic partitioning estimators have also been studied in the context of estimating piecewise polynomial densities; see Willett and Nowak (2007). This current paper focusses on the regression setting and follows this line of work of studying optimal decision trees, proving an oracle risk bound and then investigating implications for certain function classes of interest. The optimal decision trees we study in this paper are computable in time polynomial in the sample size.

In particular, in this paper, we study two decision tree methods for estimating regression functions in general dimensions in the context of estimating some nonsmooth function classes of recent interest. We focus on the fixed lattice design case like in Donoho (1997). The first method is an optimal dyadic regression tree and is exactly the same as Dyadic CART in Donoho (1997) when the dimension is 2. The second method is an Optimal Regression Tree (ORT), very much in the sense of Bertsimas and Dunn (2017), applied to fixed lattice design regression. Here the estimator is computed by optimizing a penalized least squares criterion over the set of all — not just dyadic — decision trees. We make the crucial observation that this estimator can be computed by a dynamic programming approach when the design points fall on a lattice. Thus, for instance, one does not need to resort to mixed integer programming and this dynamic program has computational complexity polynomial in the sample size. This observation may be known to the experts but we are unaware of an exact reference. Like in Donoho (1997) we show it is possible to prove an oracle risk bound (see Theorem 2.1) for both of our optimal decision tree estimators. We then apply this oracle risk bound to three function classes of recent interest by employing approximation theoretic inequalities and show that these optimal decision trees have excellent adaptive and worst case performance.

Overall in this paper, we revisit the classical idea of recursive partitioning in the context of finding answers to several unsolved questions about some classes of functions of recent interest in the nonparametric regression literature. In the course of doing so, we have come up with as well as brought forward several interesting ideas from different areas relevant for the study of regression trees such as dynamic programming, computational geometry and discrete Sobolev type inequalities for vector / matrix approximation. We believe that the main novel aspect of the current work is to recognize, prove and point out — by an amalgamation of these ideas — that optimal regression trees often provide a better alternative to the state of the art convex optimization methods in the sense they are simultaneously (near-) minimax rate optimal, adaptive to the complexity of the underlying signal (under fewer assumptions) and computationally more efficient for some classes of functions of recent interest. To the best of our knowledge, our paper is the first one among a series of recent

works that shows the efficacy of computationally efficient optimal regression tree estimators in these particular nonparametric regression problems. We now describe the function classes we consider in this paper and briefly outline our results and contributions.

- **Piecewise Polynomial Functions**: We address the problem of estimating multivariate functions that are (or close to) *piecewise polynomial* of some fixed degree on some unknown partition of the domain into axis aligned rectangles. This includes function classes such as piecewise constant/linear/quadratic etc. on axis aligned rectangles. An oracle, who knows the true rectangular partition, i.e the number of axis aligned rectangles and their arrangement, can just perform least squares separately for data falling within each rectangle. This oracle estimator provides a benchmark for adaptively optimal performance. The main question of interest to us is how to construct an estimator which is efficiently computable and attains risk as close as possible to the risk of this oracle estimator. To the best of our knowledge, this question has not been answered in multivariate settings. In this paper, we propose that our optimal regression tree (ORT) estimator solves this question to a considerable extent. Section 3 describes all of our results under this topic. It is worthwhile to mention here that we also focus on cases where the true rectangular partition does not correspond to any decision tree (see Figure 1.2.2) which necessarily has a hierarchical structure. We call such partitions nonhierarchical. Even for such nonhierarchical partitions, we make the case that ORT continues to perform well (see our results in Section 3.2.1). We are not aware of nonhierarchical partitions being studied before in the literature. Here our proof technique uses results from computational geometry which relate the size of any given (possibly nonhierarchical) rectangular partition to that of the minimal hierarchical partition refining it.
- **Multivariate Bounded Variation Functions**: Consider the function class whose total variation (defined later in Section 4) is bounded by some number. This is a classical function class for nonparametric regression since it contains functions which demonstrate spatially heterogenous smoothness; see Section 6.2 in Tibshirani (2015) and references therein. Perhaps, the most natural estimator for this class of functions is what is called the Total Variation Denoising (TVD) estimator. The two dimensional version of this estimator is also very popularly used for image denoising; see Rudin et al. (1992). It is known that a well tuned TVD estimator is minimax rate optimal for this class in all dimensions; see Hütter and Rigollet (2016) and Sadhanala et al. (2016). Also, in the univariate case, it is known that the TVD estimator adapts to piecewise constant functions and attains a near oracle risk with parametric rate of convergence; see Guntuboyina et al. (2020) and references therein. However, even in two dimensions, the TVD estimator provably cannot attain the near parametric rate of convergence for piecewise constant truths. This is a result (Theorem 2.3) in a previous article by the same authors Chatterjee and Goswami (2019).
  It would be desirable for an estimator to attain the minimax rate among bounded variation functions as well as retain the near parametric rate of convergence for piecewise constant truths in multivariate settings. Our contribution here is to establish that

Dyadic CART enjoys these two desired properties in all dimensions. We also show that the Dyadic CART adapts to the intrinsic dimensionality of the function in a particular sense. Therorem 4.2 is our main result under this topic. Our proof technique for Theorem 4.2 involves a recursive partitioning strategy to approximate any given bounded variation function by a piecewise constant function (see Proposition C.5). We prove an inequality which can be thought of as the discrete version of the classical Gagliardo-Sobolev-Nirenberg inequality (see Proposition C.7) which plays a key role in the proof.

As far as we are aware, Dyadic CART has not been investigated before in the context of estimating bounded variation functions. Coupled with the fact that Dyadic CART can be computed in time linear in the sample size, our results put forth the Dyadic CART estimator as a fast and viable option for estimating bounded variation functions.

- **Univariate Bounded Variation Functions of higher order**: Higher order versions of the space of bounded variation functions has also been considered in nonparametric regression, albeit mostly in the univariate case. One can consider the univariate function class of all r times (weakly) differentiable functions, whose r th derivative is of bounded variation. A seminal result of Donoho and Johnstone (1998) shows that a wavelet threshholding estimator attains the minimax rate in this problem. Locally adaptive regression splines, proposed by Mammen and van de Geer (1997), is also known to achieve the minimax rate in this problem. Recently, Trend Filtering, proposed by Kim et al. (2009), has proved to be a popular nonparametric regression method. Trend Filtering is very closely related to locally adaptive regression splines and is also minimax rate optimal over the space of higher order bounded variation functions; see Tibshirani (2014) and references therein. Moreover, it is known that Trend Filtering adapts to functions which are piecewise polynomials with regularity at the knots. If the number of pieces is not too large and the length of the pieces is not too small, a well tuned Trend Filtering estimator can attain near parametric risk as shown in Guntuboyina et al. (2020).

  Our main contribution here is to show that the univariate Dyadic CART estimator is also minimax rate optimal in this problem and enjoys near parametric rate of convergence for piecewise polynomials; see Theorem 5.1 and Theorem 5.2. Moreover, we show that Dyadic CART requires less regularity assumptions on the true function than what Trend Filtering requires for the near parametric rate of convergence to hold. Theorem 5.2 follows directly from a combination of our oracle risk bound and a result about refining an arbitrary (possibly non dyadic) univariate partition to a dyadic one (see Lemma C.3). Our proof technique for Theorem 5.1 again involves a recursive partitioning strategy to approximate any given higher order bounded variation function by a piecewise polynomial function (see Proposition C.9). We prove an inequality (see Lemma C.10) quantifying the error of approximating a higher order bounded variation function by a single polynomial which plays a key role in the proof. Again, as far as we are aware, Dyadic CART has not been investigated before in the context of estimating univariate higher order bounded variation functions. Coupled with the fact that Dyadic CART is computable in time linear in the sample size, our

results again provide a fast and viable alternative for estimating univariate higher order bounded variation functions.

The oracle risk bound in Theorem 2.1 which holds for the optimal decision trees studied in this paper may imply near optimal results for other function classes as well. In Section B, we mention some consequences of our oracle risk bounds for shape constrained function classes. We then describe a version of our estimators which can be implemented for arbitrary data with random design and also discuss an extension of our results for dependent noise.

1.1. *Problem Setting and Definitions.* Let us denote the d dimensional lattice with N points by $L_{d,n} := \{1, \ldots, n\}^d$ where $N = n^d$. Throughout this paper we will consider the standard fixed design setting where we treat the N design points as fixed and located on the d dimensional grid/lattice $L_{d,n}$. One may think of the design points embedded in $[0,1]^d$ and of the form $\frac{1}{n}(i_1, \ldots, i_d)$ where $(i_1, \ldots, i_d) \in L_{d,n}$. This lattice design is quite commonly used for theoretical studies in multidimensional nonparametric function estimation (see, e.g. Nemirovski (2000)). The lattice design is also the natural setting for certain applications such as image denoising, matrix/tensor estimation. All our results will be for the lattice design setting. In Section B, we make some observations and comments about possible extensions to the random design case.

Letting θ^* denote the evaluation on the grid of the underlying regression function f, our observation model becomes $y = \theta^* + \sigma Z$ where y, θ^*, Z are real valued functions on $L_{d,n}$ and hence are d dimensional arrays. Furthermore, Z is a noise array consisting of independent standard Gaussian entries and $\sigma > 0$ is an unknown standard deviation of the noise entries. For an estimator $\widehat{\theta}$, we will evaluate its performance by the usual fixed design expected mean squared error

$$\mathrm{MSE}(\widehat{\theta}, \theta^*) := \frac{1}{N} \mathbb{E}_{\theta^*} \|\widehat{\theta} - \theta^*\|^2.$$

Here $\|.\|$ refers to the usual Euclidean norm of an array where we treat an array as a vector in $\mathbb{R}^N$.

Let us define the interval of positive integers $[a, b] = \{i \in \mathbb{Z}_+ : a \leq i \leq b\}$ where $\mathbb{Z}_+$ denotes the set of positive integers. For a positive integer n we also denote the set $[1, n]$ by just $[n]$. A subset $R \subset L_{d,n}$ is called an *axis aligned rectangle* if R is a product of intervals, i.e. $R = \prod_{i=1}^d [a_i, b_i]$. Henceforth, we will just use the word rectangle to denote an axis aligned rectangle. Let us define a *rectangular partition* of $L_{d,n}$ to be a set of rectangles $\mathcal{R}$ such that (a) the rectangles in $\mathcal{R}$ are pairwise disjoint and (b) $\cup_{R \in \mathcal{R}} R = L_{d,n}$.

Recall that a multivariate polynomial of degree at most $r \geq 0$ is a finite linear combination of the monomials $\Pi_{i=1}^d (x_i)^{r_i}$ satisfying $\sum_{i=1}^d r_i \leq r$. It is thus clear that they form a linear space of dimension $K_{r,d} := \binom{r+d-1}{d-1}$. Let us now define the set of discrete multivariate

polynomial arrays as

$$\mathcal{F}^{(r)}_{d,n} = \big\{\theta \in \mathbb{R}^{L_{d,n}} : \theta(i_1/n, \ldots, i_d/n) = f(i_1/n, \ldots, i_d/n)\ \forall (i_1, \ldots, i_d) \in [n]^d \text{ for some polynomial } f \text{ of degree at most } r\big\}.$$

For a given rectangle $R \subset L_{d,n}$ and any $\theta \in \mathbb{R}^{L_{d,n}}$ let us denote the array obtained by restricting θ to R by θ_R. We say that θ is a degree r polynomial on the rectangle R if $\theta_R = \alpha_R$ for some $\alpha \in \mathcal{F}^{(r)}_{d,n}$.

For a given array $\theta \in \mathbb{R}^{L_{d,n}}$, let $k^{(r)}(\theta)$ *denote the smallest positive integer* k *such that a set of* k *rectangles* $R_1, \ldots, R_k$ *form a rectangular partition of* $L_{d,n}$ *and the restricted array* θ_{R_i} *is a degree* r *polynomial for all* $1 \le i \le k$. In other words, $k^{(r)}(\theta)$ is the cardinality of the minimal rectangular partition of $L_{d,n}$ such that θ is piecewise polynomial of degree r on the partition.

1.2. *Description of Estimators.* The estimators we consider in this manuscript compute a data dependent decision tree (which is globally optimal in a certain sense) and then fit polynomials within each cell/rectangle of the decision tree. As mentioned before, computing decision trees greedily and then fitting a constant value within each cell of the decision tree has a long history and is what the usual CART does. Fitting polynomials on such greedily grown decision trees is a natural extension of CART and has also been proposed in the literature; see Chaudhuri et al. (1994). The main difference between these estimators and our estimators is that our decision trees are computed as a global optimizer over the set of all decision trees. In particular, they are not grown greedily and there is no stopping rule that is required. The ideas here are mainly inspired by Donoho (1997). We now define our estimators precisely.

Recall the definition of $k^{(r)}(\theta)$. A natural estimator which fits piecewise polynomial functions of degree $r \ge 0$ on axis aligned rectangles is the following fully penalized LSE of order r:

$$\widehat{\theta}^{(r)}_{\text{all},\lambda} := \operatorname*{argmin}_{\theta \in \mathbb{R}^{L_{d,n}}} \big(\|y - \theta\|^2 + \lambda k^{(r)}(\theta)\big).$$

Let us denote the set of all rectangular partitions of $L_{d,n}$ as $\mathcal{P}_{\text{all},d,n}$. For each rectangular partition $\Pi \in \mathcal{P}_{\text{all},d,n}$ and each nonnegative integer r, let the (linear) subspace $S^{(r)}(\Pi)$ comprise all arrays which are degree r polynomial on each of the rectangles constituting Π. For a generic subspace $S \subset \mathbb{R}^N$ let us denote its dimension by $Dim(S)$ and the associated orthogonal projection matrix by O_S. Clearly the dimension of the subspace $S^{(r)}(\Pi)$ is $K_{r,d}|\Pi|$ where $|\Pi|$ is the cardinality of the partition. Now note that we can also write $\widehat{\theta}^{(r)}_{\text{all},\lambda} = O_{S^{(r)}(\widehat{\Pi}(\lambda))} y$ where $\widehat{\Pi}(\lambda)$ is a data dependent partition defined as

$$\widehat{\Pi}(\lambda) = \operatorname*{argmin}_{\Pi : \Pi \in \mathcal{P}_{\text{all},d,n}} \big(\|y - O_{S^{(r)}(\Pi)} y\|^2 + \lambda |\Pi|\big) \tag{1.1}$$

Thus, computing $\widehat{\theta}^{(r)}_{\lambda,\text{all}}$ really involves optimizing over all rectangular partitions $\Pi \in \mathcal{P}_{\text{all},d,n}$. Therefore, one may anticipate that the major roadblock in using this estimator would be computation. For any fixed d, the cardinality of $\mathcal{P}_{\text{all},d,n}$ is at least stretched-exponential in N. Thus, a brute force method is infeasible. However, for $d = 1$, a rectangular partition is a set of contiguous blocks of intervals which has enough structure so that a dynamic programming approach is amenable. The set of all multivariate rectangular partitions is a more complicated object and the corresponding computation is likely to be provably hard. This is where the idea of Donoho (1997) comes in who considers the Dyadic CART estimator (for $r = 0$ and $d = 2$) for fitting piecewise constant functions. As we will now explain, it turns out that if we constrain the optimization in (1.1) to optimize over special subclasses of rectangular partitions of $L_{d,n}$, a dynamic programming approach again becomes tractable. The Dyadic CART estimator is one such constrained version of the optimization problem in (1.1). We now precisely define these subclasses of rectangular partitions.

1.2.1. *Description of Dyadic CART of order $r \geq 0$.* Let us consider a generic discrete interval $[a, b]$. We define a *dyadic split* of the interval to be a split of the interval $[a, b]$ into two equal intervals. To be concrete, the interval $[a, b]$ is split into the intervals $[a, a - 1 + \lceil (b-a+1)/2 \rceil]$ and $[a + \lceil (b-a+1)/2 \rceil, b]$. Now consider a generic rectangle $R = \prod_{i=1}^{d} [a_i, b_i]$. A *dyadic split* of the rectangle R involves the choice of a coordinate $1 \leq j \leq d$ to be split and then the j-th interval in the product defining the rectangle R undergoes a dyadic split. Thus, a dyadic split of R produces two sub rectangles R_1 and R_2 where $R_2 = R \cap R_1^c$ and R_1 is of the following form for some $j \in [d]$,

$$R_1 = \prod_{i=1}^{j-1} [a_i, b_i] \times [a_j, a_j - 1 + \lceil (b_j - a_j + 1)/2 \rceil] \times \prod_{i=j+1}^{d} [a_i, b_i].$$

Starting from the trivial partition which is just $L_{d,n}$ itself, we can create a refined partition by dyadically splitting $L_{d,n}$. This will result in a partition of $L_{d,n}$ into two rectangles. We can now keep on dividing recursively, generating new partitions. In general, if at some stage we have the partition $\Pi = (R_1, \ldots, R_k)$, we can choose any of the rectangles R_i and dyadically split it to get a refinement of Π with $k+1$ nonempty rectangles. *A recursive dyadic partition* (RDP) is any partition reachable by such successive dyadic splitting. Let us denote the set of all recursive dyadic partitions of $L_{d,n}$ as $\mathcal{P}_{\text{rdp},d,n}$. Indeed, a natural way of encoding any RDP of $L_{d,n}$ is by a binary tree where each nonleaf node is labeled by an integer in $[d]$. This labeling corresponds to the choice of the coordinate that was used for the split.

We can now consider a constrained version of $\widehat{\theta}^{(r)}_{\text{all},\lambda}$ which only optimizes over $\mathcal{P}_{\text{rdp},d,n}$ instead of optimizing over $\mathcal{P}_{\text{all},d,n}$. Let us define $\widehat{\theta}^{(r)}_{\text{rdp},\lambda} = O_{S^{(r)}(\widehat{\Pi}_{\text{rdp}}(\lambda))} y$ where $\widehat{\Pi}_{\text{rdp}}(\lambda)$ is a data dependent partition defined as

$$\widehat{\Pi}_{\text{rdp}}(\lambda) = \operatorname*{argmin}_{\Pi : \Pi \in \mathcal{P}_{\text{rdp},d,n}} \big(\|y - O_{S^{(r)}(\Pi)}\|^2 + \lambda |\Pi| \big).$$

The estimator $\widehat{\theta}^{(r)}_{\mathrm{rdp},\lambda}$ is precisely the Dyadic CART estimator which was proposed in Donoho (1997) in the case when $d = 2$ and $r = 0$. The author studied this estimator for estimating anisotropic smooth functions of two variables which exhibit different degrees of smoothness in the two variables. However, to the best of our knowledge, the risk properties of the Dyadic CART estimator (for $r = 0$) has not been examined in the context of estimating nonsmooth function classes such as piecewise constant and bounded variation functions. For $r \geq 1$, the above estimator appears to not have been proposed and studied in the literature before. We call the estimator $\widehat{\theta}^{(r)}_{\mathrm{rdp},\lambda}$ as *Dyadic CART of order* r.

1.2.2. *Description of ORT of order* $r \geq 0$. For our purposes, we would need to consider a larger class of partitions than $\mathcal{P}_{\mathrm{rdp},d,n}$. To generate a RDP, for each rectangle we choose a dimension to split and then split at the midpoint. Instead of splitting at the midpoint, it is natural to allow the split to be at an arbitrary position. To that end, we define a *hierarchical split* of the interval to be a split of the interval $[a, b]$ into two intervals, but not necessarily equal sized. To be concrete, the interval $[a, b]$ is split into the intervals $[a, \ell]$ and $[\ell + 1, b]$ for some $a \leq \ell \leq b$. Now consider a generic rectangle $R = \prod_{i=1}^{d}[a_i, b_i]$. A *hierarchical split* of the rectangle R involves the choice of a coordinate $1 \leq j \leq d$ to be split and then the j-th interval in the product defining the rectangle R undergoes a hierarchical split. Thus, a hierarchical split of R produces two sub rectangles R_1 and R_2 where $R_2 = R \cap R_1^c$ and R_1 is of the following form for some $1 \leq j \leq d$ and $a_j \leq \ell \leq b_j$,

$$R_1 = \prod_{i=1}^{j-1}[a_i, b_i] \times [a_j, \ell] \times \prod_{i=j+1}^{d}[a_i, b_i].$$

Starting from the trivial partition $L_{d,n}$ itself, we can now generate partitions by splitting $L_{d,n}$ hierarchically. Again, in general if at some stage we obtain the partition $\Pi = (R_1, \ldots, R_k)$, we can choose any of the rectangles R_i and split it hierarchically to obtain $k + 1$ nonempty rectangles now. *A hierarchical partition* is any partition reachable by such hierarchical splits. We denote the set of all hierarchical partitions of $L_{d,n}$ as $\mathcal{P}_{\mathrm{hier},d,n}$. Note that a hierarchical partition is in one to one correspondence with decision trees and thus, $\mathcal{P}_{\mathrm{hier},d,n}$ can be thought of as the set of all decision trees.

Clearly,

$$\mathcal{P}_{\mathrm{rdp},d,n} \subset \mathcal{P}_{\mathrm{hier},d,n} \subset \mathcal{P}_{\mathrm{all},d,n}.$$

In fact, the inclusions are strict as shown in Figure 1. In particular, there exist partitions which are not hierarchical.

We can now consider another constrained version of $\widehat{\theta}^{(r)}_{\mathrm{all},\lambda}$ which optimizes only over $\mathcal{P}_{\mathrm{hier},d,n}$. Let us define $\widehat{\theta}^{(r)}_{\mathrm{hier},\lambda} = O_{S^{(r)}(\widehat{\Pi}_{\mathrm{hier}}(\lambda))}y$ where $\widehat{\Pi}_{\mathrm{hier}}(\lambda)$ is a data dependent partition defined as

$$\widehat{\Pi}_{\mathrm{hier}}(\lambda) = \operatorname*{argmin}_{\Pi:\Pi\in\mathcal{P}_{\mathrm{hier},d,n}} \big(\|y - O_{S^{(r)}(\Pi)}y\|^2 + \lambda|\Pi|\big).$$

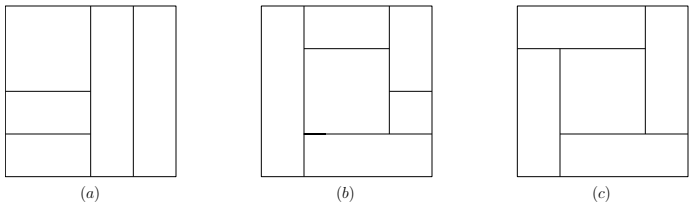

FIG 1. *Figure (a) is an example of a recursive dyadic partition of the square. Figure (b) is nondyadic but is a hierarchical partition. Figure (c) is an example of a nonhierarchical partition. An easy way to see this is that there is no split from top to bottom or left to right.*

Although this is a natural extension of Dyadic CART, we are unable to pinpoint an exact reference where this estimator has been explicitly proposed or studied in the statistics literature. The above optimization problem is an analog of the optimal decision tree problem laid out in Bertsimas and Dunn (2017). The difference is that Bertsimas and Dunn (2017) is considering classification whereas we are considering fixed lattice design regression. Note that the above optimization problem is different from the usual pruning of a tree that is done at the second stage of CART. Pruning can only result in subtrees of the full tree obtained in the first stage whereas the above optimization is over all rectangular partitions $\Pi \in \mathcal{P}_{\text{hier},d,n}$. We name the estimator $\widehat{\theta}^{(r)}_{\lambda,\text{hier}}$ as *Optimal Regression Tree* (ORT) of order r.

1.3. *Both Dyadic CART and ORT of all orders are efficiently computable.* The crucial fact about $\widehat{\theta}^{(r)}_{\text{rdp},\lambda}$ and $\widehat{\theta}^{(r)}_{\text{hier},\lambda}$ is that they can be computed efficiently and exactly using dynamic programming approaches. A dynamic program algorithm to compute $\widehat{\Pi}_{\text{rdp},\lambda}$ for $d = 2$ and $r = 0$ was shown in Donoho (1997). This algorithm is extremely fast and can be computed in $O(N)$ (linear in sample size) time. The basic idea there can actually be extended to compute both Dyadic CART and ORT for any fixed r, d with computational complexity given in the next lemma. The proof is given in Section C (in the supplementary file,).

LEMMA 1.1. *There exists an absolute constant $C > 0$ such that the computational complexity, i.e. the number of elementary operations involved in the computation of ORT is bounded by:*

$$\begin{cases} CN^2\, nd, & \text{for } r = 0 \\ CN^2\,(nd + d^{3r}) & \text{for } r \geq 1 \end{cases}$$

Similarly, the computational complexity of Dyadic CART is bounded by:

$$\begin{cases} C2^d\, Nd, & \text{for } r = 0 \\ C2^d\, N\, d^{3r} & \text{for } r \geq 1. \end{cases}$$

REMARK 1.1. *Since the proxy for the sample size $N \geq 2^d$ as soon as $n \geq 2$, it does not make sense to think of d as large when reading the above computational complexity. The lattice design setting is really meaningful when d is low to moderate and fixed and the number*

of samples per dimension n is growing to ∞. Thus, one should look at the dependence of the computational complexity on N and treat the factors depending on d as constant.

REMARK 1.2. *Even for $d = 1$, the brute force computation time is exponential in N as the total number of hierarchical partitions is exponential in N.*

The rest of the paper is organized as follows. In Section 2 we state our oracle risk bound for Dyadic CART and ORT of all orders. Section 3 describes applications of the oracle risk bound for ORT to multivariate piecewise polynomials. In Sections 4 and 5 we state applications of the oracle risk bound for Dyadic CART to bounded variation functions in general dimensions and univariate bounded variation function classes of all orders respectively. In Section A of the supplementary file, we describe our simulation studies and in Section B, we discuss the main contributions of this paper vis a vis the relevant methodologies present in the literature as well as discuss some possible extensions of the scope of our results. Section C in the supplementary file contains all the proofs of our results. In Section D in the same file we state and prove some auxiliary results that we use for proving our main results in the paper.

Acknowledgements: This research was supported by a NSF grant and an IDEX grant from Paris-Saclay. S.G.'s research was carried out in part as a member of the Infosys-Chandrasekharan virtual center for Random Geometry, supported by a grant from the Infosys Foundation. We thank the anonymous referees for their numerous helpful remarks and suggestions on an earlier manuscript of the paper. We also thank Adityanand Guntuboyina for many helpful comments. The project started when SG was a postdoctoral fellow at the Institut des Hautes Études Scientifiques (IHES).

2. Oracle risk bounds for Dyadic CART and ORT. In this section we describe an oracle risk bound. We have to set up some notations and terminology first. Let $\mathcal{S}$ be any finite collection of subspaces of $\mathbb{R}^N$. Recall that for a generic subspace $S \in \mathcal{S}$, we denote its dimension by $Dim(S)$. For any given $\theta \in \mathbb{R}^N$ let us define

$$k_{\mathcal{S}}(\theta) = \min\{Dim(S) : S \in \mathcal{S}, \theta \in S\} \tag{2.1}$$

where we adopt the convention that the infimum of an empty set is ∞.

For any $\theta \in \mathbb{R}^N$, the number $k_{\mathcal{S}}(\theta)$ can be thought of as describing the complexity of θ with respect to the collection of subspaces $\mathcal{S}$. Recall the definition of the nested classes of rectangular partitions $\mathcal{P}_{\mathrm{rdp},d,n} \subset \mathcal{P}_{\mathrm{hier},d,n} \subset \mathcal{P}_{\mathrm{all},d,n}$. Also recall that the subspace $S^{(r)}(\Pi)$ denotes all arrays which are degree r polynomial on each of the rectangles constituting Π. For any integer $r \geq 0$, these classes of partitions induce their respective collection of subspaces of $\mathbb{R}^N$ defined as follows:

$$\mathcal{S}_a^{(r)} = \{S^{(r)}(\Pi) : \Pi \in \mathcal{P}_{a,d,n}\}$$

where $a \in \{\text{rdp}, \text{hier}, \text{all}\}$. For any $\theta \in \mathbb{R}^{L_{d,n}}$ and any integer $r \geq 0$ let us observe its complexity with respect to the collection of subspaces $S_a^{(r)}$ is

$$k_{\mathcal{S}_a^{(r)}}(\theta) = k_a^{(r)}(\theta)$$

where again $a \in \{\text{rdp}, \text{hier}, \text{all}\}$. Here $k_{\text{all}}^{(r)}(\theta^*)$ is the same as $k^{(r)}(\theta^*)$ defined earlier and we use both notations interchangeably.

It is now clear that for any $\theta \in \mathbb{R}^N$ we have

$$k_{\text{all}}^{(r)}(\theta) \leq k_{\text{hier}}^{(r)}(\theta) \leq k_{\text{rdp}}^{(r)}(\theta). \tag{2.2}$$

We are now ready to state an oracle risk bound for all the three estimators $\widehat{\theta}_{\text{all},\lambda}^{(r)}, \widehat{\theta}_{\text{rdp},\lambda}^{(r)}$ and $\widehat{\theta}_{\text{hier},\lambda}^{(r)}$. The theorem is proved in Section C.

THEOREM 2.1. *Fix any integer $r \geq 0$ and recall that $K_{r,d} = \binom{r+d-1}{d-1}$ was defined earlier. There exists an absolute constant $C > 0$ such that for any $0 < \delta < 1$ if we set $\lambda \geq CK_{r,d}\frac{\sigma^2 \log N}{\delta}$, then we have the following risk bounds for $a \in \{\text{rdp}, \text{hier}, \text{all}\}$,*

$$\mathbb{E}\|\widehat{\theta}_{a,\lambda}^{(r)} - \theta^*\|^2 \leq \inf_{\theta \in \mathbb{R}^N} \big[\frac{(1+\delta)}{(1-\delta)}\,\|\theta - \theta^*\|^2 + \frac{\lambda}{1-\delta}\,k_a^{(r)}(\theta)\big] + C\frac{\sigma^2}{\delta\,(1-\delta)}$$

REMARK 2.1. *Operationally, to derive risk bounds for Dyadic CART or ORT for some function class, Theorem 2.1 behooves us to use approximation theoretic arguments. To be more precise, for a given generic θ^* in the function class, one needs to understand what is the approximation error in the Euclidean sense, if the approximator θ is constrained to satisfy $k_{\text{rdp}}(\theta) = k$ or $k_{\text{hier}}(\theta) = k$ for any given integer k. One of the technical contributions of this paper lies in addressing this approximation theoretic question for the three classes of functions considered in this paper.*

Several other remarks about the above theorem are presented in Section B.1 (in the supplementary file) due to space considerations.

3. Results for Multivariate Piecewise Polynomial Functions.

3.1. *The main question.* For a given underlying truth θ^*, the *oracle estimator* $\widehat{\theta}_{(\text{oracle})}^{(r)}$ — which *knows* the minimal rectangular partition $(R_1, \ldots, R_k)$ of θ^* *exactly* — has a simple form. In words, within each rectangle R_i, it estimates $\theta^*_{R_i}$ by the best fitting r-th degree polynomial in the least squares sense. It is not hard to check that

$$\text{MSE}(\widehat{\theta}_{(\text{oracle})}^{(r)}, \theta^*) \leq K_{r,d}\,\sigma^2\frac{k_{\text{all}}^{(r)}(\theta^*)}{N}.$$

Thus, for any fixed d and r, the MSE of the oracle estimator scales like the *number of pieces* $k_{\text{all}}^{(r)}(\theta^*)$ divided by the sample size N which is precisely the parametric rate of convergence. Furthermore, we can show the following minimax lower bound holds.

LEMMA 3.1. *Fix any positive integers n, d. Fix any integer k such that $3d \leq k \leq N = n^d$ and let $\Theta_{k,d,n} := \{\theta \in \mathbb{R}^{L_{d,n}} : k_{\text{hier}}^{(0)}(\theta) \leq k\}$. There exists a universal constant C such that the following inequality holds:*

$$\inf_{\widetilde{\theta}} \sup_{\theta \in \Theta_{k,d,n}} \mathbb{E}\|\widetilde{\theta} - \theta\|^2 \geq C\, \sigma^2 k \log \frac{\mathrm{e}N}{k}.$$

Here the infimum is over all estimators $\widetilde{\theta}$ which are measurable functions of the data array y.

REMARK 3.1. *For any $r \geq 1$, since $k_{\text{hier}}^{(0)}(\theta) \geq k_{\text{all}}^{(0)}(\theta) \geq k_{\text{all}}^{(r)}(\theta)$ the same minimax lower bound also holds for the parameter space $\{\theta \in \mathbb{R}^{L_{d,n}} : k_{\text{all}}^{(r)}(\theta) \leq k\}$.*

The above minimax lower bound shows that any estimator must incur MSE (in the worst case) which is the oracle MSE multiplied by an extra $\log(eN/k)$ factor. In particular, if $k = o(N)$, which is the interesting regime, the extra $\log N$ factor is inevitable. We call this $O\big(\frac{k}{N} \log(eN/k)\big)$ rate the minimax rate from here on.

We provide the proof of Lemma 3.1 in Section C.8. We now ask the following question for every fixed dimension d and degree r.

Q: Does there exist an estimator which

- *attains the minimax rate* MSE *scaling like $O\big(\sigma^2 \frac{k_{\text{all}}^{(r)}(\theta^*)}{N} \log N\big)$ for all θ^* adaptively, and*
- *is possible to compute in polynomial time in the sample size $N = n^d$?*

To the best of our knowledge, the above question relating to computationally efficient minimax adaptive estimation of piecewise polynomial functions in multivariate settings, even for piecewise constant functions in the planar case (i.e. $r = 0, d = 2$), has not been rigorously answered in the statistics literature. The fully penalized least squares estimator $\widehat{\theta}_{\text{all},\lambda}^{(r)}$ is naturally suited for our purpose but is likely to be computationally infeasible. The goal of this section is to show that

- In the two dimensional setting, i.e. $d = 2$, the ORT estimator attains the minimax MSE rate adaptively for any truth θ^*. The ORT attains this minimax rate even if the true underlying rectangular partition is not hierarchical. In particular, we show that the ORT incurs the oracle MSE with the exponent of $\log N$ equalling 1 thus matching the minimax lower bound in Lemma 3.1 up to constant factors.

- When $d > 2$, as long as the true underlying rectangular partition satisfies natural regularity conditions such as being hierarchical or fat (defined later in this section), the ORT estimator continues to attain this minimax rate.

We prove these results by combining Theorem 2.1 with existing results in computational geometry. To the best of our knowledge, our results in this section are the first of their type. We also give a review of what is known in the univariate case (when $d = 1$) in Section B.2 (in the supplementary file).

3.2. *Our Results for ORT.* In the remainder of the paper the constant involved in $O(\cdot)$ may depend on r and d unless specifically mentioned otherwise. Also to lighten the notation, we use $\widetilde{O}(\cdot)$ for $O(\cdot)\text{poly}(\log N)$. Recall that $K_{r,d}$ is the dimension of the subspace of d dimensional polynomials with degree at most $r \geq 0$. An immediate corollary of Theorem 2.1 is the following.

COROLLARY 3.2. *There exists an absolute constant $C > 0$ such that by setting $\lambda = CK_{r,d}\,\sigma^2\,\log N$ we have the following risk bound,*

$$MSE(\widehat{\theta}^{(r)}_{\text{hier},\lambda}, \theta^*) \leq \frac{C\,K_{r,d}\;\sigma^2\;\log N}{N}\,k^{(r)}_{\text{hier}}(\theta^*) + \frac{C\,\sigma^2}{N}.$$

Let us discuss some implications of the above corollary. For ORT of order $r \geq 0$, a risk bound scaling like $O(\frac{k^{(r)}_{\text{hier}}(\theta^*)}{N}\log N)$ is guaranteed for all θ^*. Thus, for instance, if the true θ^* is piecewise constant/linear on some arbitrary unknown hierarchical partition of $L_{d,n}$, the corresponding ORT estimator of order $0, 1$ respectively achieves the (near) minimax risk $O(\frac{k^{(r)}_{\text{all}}(\theta^*)}{N}\log N)$. Although this result is an immediate implication of Theorem 2.1, this is the first such risk guarantee established for a computationally efficient decision tree estimator in general dimensions as far as we are aware of.

At this point, let us recall that our target is to achieve the ideal upper bound $\widetilde{O}(\frac{k^{(r)}_{\text{all}}(\theta^*)}{N})$ to the MSE for all θ^* which is attained by the fully penalized LSE. However, it is perhaps not efficiently computable. The best upper bound to the MSE we can get for a computationally efficient estimator is $\widetilde{O}(\frac{k^{(r)}_{\text{hier}}(\theta^*)}{N})$ which is attained by the ORT estimator.

A natural question that arises at this point is how much worse is the upper bound for ORT than the upper bound for the fully penalized LS estimator given in Theorem 2.1. Equivalently, we know that $k^{(r)}_{\text{all}}(\theta^*) \leq k^{(r)}_{\text{hier}}(\theta^*)$ in general, but how large can the gap be? There definitely exist partitions which are not hierarchical, i.e. that is $\mathcal{P}_{\text{hier},d,n}$ is a strict subset of $\mathcal{P}_{\text{all},d,n}$ as shown in Figure 1.

In the next section we explore general and possibly nonhierarchical partitions of $L_{d,n}$ and state several results which basically imply that ORT incurs MSE at most a constant fac-

tor more than the ideal fully penalized LSE for several natural instances of rectangular partitions.

3.2.1. *Arbitrary partitions.* The risk bound for ORT in Theorem 2.1 is in terms of $k_{\text{hier}}(\theta^*)$. We would like to convert it into a risk bound involving $k_{\text{all}}(\theta^*)$. A natural way of doing this would be to refine an arbitrary partition into a hierarchical partition and then count the number of extra rectangular pieces that arises as a result of this refinement. This begs the following question of a combinatorial flavour.

Can an arbitrary partition of $L_{d,n}$ be refined into a hierarchical partition without increasing the number of rectangles too much?.

Fortunately, the above question has been studied a fair bit in the computational/combinatorial geometry literature under the name of *binary space partitions*. A binary space partition (BSP) is a recursive partitioning scheme for a set of objects in space. The goal is to partition the space recursively until each smaller space contains only one/few of the original objects. The main questions of interest are, given the set of objects, the minimal cardinality of the optimal partition and an efficient algorithm to compute it. A nice survey of this area, explaining the central questions and an overview of known results can be found in Tóth (2005). We will now leverage some existing results in this area which would yield corresponding risk bounds with the help of Theorem 2.1.

For $d = 2$, it turns out that any rectangular partition can be refined into a hierarchical one where the number of rectangular pieces at most doubles. The following proposition is due to Berman et al. (2002) and states this fact.

PROPOSITION 3.3 (Berman et al. (2002)). *Given any partition $\Pi \in \mathcal{P}_{\text{all},2,n}$ there exists a refinement $\widetilde{\Pi} \in \mathcal{P}_{\text{hier},2,n}$ such that $|\widetilde{\Pi}| \leq 2|\Pi|$. As a consequence, for any matrix $\theta \in \mathbb{R}^{n\times n}$ and any nonnegative integer r, we have*

$$k^{(r)}_{\text{hier}}(\theta) \leq 2k^{(r)}_{\text{all}}(\theta).$$

The above proposition applied to Theorem 2.1 immediately yields the following theorem:

THEOREM 3.4. *Let $d = 2$. There exists an absolute constant C such that by setting $\lambda = C\, K_{r,d}\ \sigma^2\ \log N$ we have the following risk bound for $\widehat{\theta}_{\text{hier},\lambda}$:*

$$\text{MSE}(\widehat{\theta}^{(r)}_{\text{hier},\lambda}, \theta^*) \leq \frac{C\, K_{r,d}\ \sigma^2\ \log N}{N}\, k^{(r)}_{\text{all}}(\theta^*) + \frac{C\,\sigma^2}{N}.$$

REMARK 3.2. *Thus, in the two dimensional setting $d = 2$,* ORT *fulfills the two objectives of computability in polynomial time and attaining the minimax risk rate adaptively for all truths θ^*. Thus, this completely solves the main question we posed in the two dimensional case. To the best of our knowledge, this is the first result of its kind in the literature.*

For dimensions higher than 2; the best result akin to Proposition 3.3 that is available is due to Hershberger et al. (2005).

PROPOSITION 3.5 (Hershberger et al. (2005)). *Let $d > 2$. Given any partition $\Pi \in \mathcal{P}_{\text{all},d,n}$ there exists a refinement $\widetilde{\Pi} \in \mathcal{P}_{\text{hier},d,n}$ such that $|\widetilde{\Pi}| \leq |\Pi|^{\frac{d+1}{3}}$. As a consequence, for any array $\theta \in \mathbb{R}^{L_{d,n}}$ and any nonnegative integer r, we have*

$$k^{(r)}_{\text{hier}}(\theta) \leq \left(k^{(r)}_{\text{all}}(\theta)\right)^{\frac{d+1}{3}}.$$

REMARK 3.3. *A matching lower bound is also given in Hershberger et al. (2005) for the case $d = 3$. Thus, to refine a rectangular partition (of k pieces) into a hierarchical one, one necessarily increases the number of rectangular pieces to $O(k^{4/3})$ in the worst case.*

The above result suggests that for arbitrary partitions in d dimensions, our current approach will not yield the near minimax rate of convergence. Nevertheless, we state our risk bound that is implied by Proposition 3.5.

THEOREM 3.6. *Let $d > 2$. There exists an absolute constant C such that by setting $\lambda \geq C\, K_{r,d}\, \sigma^2\, \log N$ we have the following risk bound for $\widehat{\theta}_{\text{hier},\lambda}$:*

$$\text{MSE}(\widehat{\theta}^{(r)}_{\text{hier},\lambda}, \theta^*) \leq \lambda \frac{\left(k^{(r)}_{\text{all}}(\theta^*)\right)^{\frac{d+1}{3}}}{N} + \frac{C\,\sigma^2}{N}.$$

Our approach of refining an arbitrary partition into a hierarchical partition does not seem to yield the $\widetilde{O}\left(\sigma^2 \frac{k^{(r)}_{\text{all}}(\theta^*)}{N}\right)$ rate of convergence for ORT in dimension higher than 2 when the truth is a piecewise polynomial function on an *arbitrary* rectangular partition. Rectangular partitions in higher dimensions could be highly complex; with some rectangles being very "skinny" in some dimensions. However, it turns out that if we rule out such anomalies, then it is still possible to attain our objective. Let us now define a class of partitions which rules out such anomalies.

Let R be a rectangle defined as $R = \Pi_{i=1}^{d}[a_i, b_i] \subset L_{d,n}$. Let the sidelengths of R be defined as $n_i = b_i - a_i + 1$ for $i \in [d]$. Define its *aspect ratio* as $\mathcal{A}(R) = \max\{\frac{n_i}{n_j} : (i,j) \in [d]^2\}$. For any $\alpha \geq 1$, let us call a rectangle α *fat* if we have $\mathcal{A}(R) \leq \alpha$. Now consider a rectangular partition $\Pi \in \mathcal{P}_{\text{all},d,n}$. We call Π an α *fat partition* if each of its constituent rectangles is α fat. Let us denote the class of α fat partitions of $L_{d,n}$ as $\mathcal{P}_{\text{fat}(\alpha),d,n}$. As before, we can now define the class of subspaces $S^{(r)}_{\text{fat}(\alpha),d,n}$ corresponding to the set of partitions $\mathcal{P}_{\text{fat}(\alpha),d,n}$. For any array θ^* and any integer $r > 0$ we can also denote

$$k^{(r)}_{\text{fat}(\alpha)}(\theta^*) = k_{S^{(r)}_{\text{fat}(\alpha),d,n}}(\theta^*).$$

An important result in the area of binary space partitions is that any fat rectangular partition of $L_{n,d}$ can be refined into a hierarchical one with the number of rectangular

pieces inflated by at most a constant factor. This is the content of the following proposition which is due to de Berg (1995).

PROPOSITION 3.7 (de Berg (1995)). *There exists a constant $C(d,\alpha) \geq 1$ depending only on d and α such that any partition $\Pi \in \mathcal{P}_{\mathrm{fat}(\alpha),d,n}$ can be refined into a hierarchical partition $\widetilde{\Pi} \in \mathcal{P}_{\mathrm{hier},d,n}$ satisfying*

$$|\widetilde{\Pi}| \leq C(d,\alpha)|\Pi|.$$

Equivalently, for any $\theta \in \mathbb{R}^{L_{n,d}}$ and any non negative integer r we have

$$k^{(r)}_{\mathrm{hier}}(\theta) \leq C(d,\alpha) k^{(r)}_{\mathrm{fat}(\alpha)}(\theta).$$

The above proposition gives rise to a risk bound for ORT in all dimensions.

THEOREM 3.8. *For any dimension d there exists an absolute constant C such that by setting $\lambda \geq C\, K_{r,d}\, \sigma^2\, \log n$ we have the following risk bound for $\widehat{\theta}_{\mathrm{hier},\lambda}$:*

$$\mathbb{E}\|\widehat{\theta}^{(r)}_{\mathrm{hier},\lambda} - \theta^*\|^2 \leq \inf_{\theta \in \mathbb{R}^{L_{n,d}}} \left(2\,\|\theta - \theta^*\|^2 + \lambda\, C(d,\alpha)\, k^{(r)}_{\mathrm{fat}(\alpha)}(\theta)\right) + C\,\sigma^2.$$

REMARK 3.4. *For any fixed dimension d, when θ^* is piecewise polynomial of degree r on a fat paritition, the above theorem implies a $O\big(\sigma^2 \frac{k^{(r)}_{\mathrm{all}}(\theta^*)}{N} \log N\big)$ bound to the MSE of the ORT estimator (of order r). Thus, for arbitrary fat partitions in any dimension, ORT attains our objective of enjoying the near minimax rate of convergence and being computationally efficient. For any fixed dimension d, this is the first result of its type that we are aware of.*

REMARK 3.5. *It should be mentioned here that the constant $C(d,\alpha)$ scales exponentially with d, at least in the construction which is due to de Berg (1995). In any case, recall that all of our results are meaningful when d is low to moderate.*

3.3. *Our Results for Dyadic CART.* In the previous section, we showed that the ORT estimator attains the desired $\widetilde{O}\big(\sigma^2 \frac{k^{(r)}_{\mathrm{all}}(\theta^*)}{N}\big)$ rate for all θ^* adaptively in dimensions $d = 1, 2$ and for all θ^* which are piecewise polynomial on a fat partition in all dimensions $d > 2$. Since the ORT is more computationally expensive than Dyadic CART, a natural question is whether there are analogous results for Dyadic CART. In this case, the relevant question is

Can an arbitrary nonhierarchical partition of $L_{d,n}$ be refined into a recursive dyadic partition without increasing the number of rectangles too much?.

When $d = 1$ or $d = 2$, we can give an argument to show there exists a recursive dyadic partition refining a given arbitrary rectangular partition with number of rectangles being multiplied by a log factor. This is the content of our next result which is proved in Section C.3.

PROPOSITION 3.9. *Given any positive integer n and given a partition $\Pi \in \mathcal{P}_{\text{all},1,n}$ with k intervals, there exists a refinement $\widetilde{\Pi} \in \mathcal{P}_{\text{rdp},1,n}$ which is a recursive dyadic partition with at most $Ck\log(en/k)$ intervals where $C > 0$ is an universal constant. Equivalently, for all $\theta \in \mathbb{R}^{L_{1,n}}$ and all non negative integers r, we have*

$$k^{(r)}_{\text{rdp}}(\theta) \leq Ck^{(r)}_{\text{all}}(\theta) \log \frac{en}{k^{(r)}_{\text{all}}(\theta)}. \tag{3.1}$$

Moreover, given any positive integer n and an arbitrary partition $\Pi \in \mathcal{P}_{\text{all},2,n}$ of $L_{2,n}$ with k rectangles there exists a refinement $\Pi' \in \mathcal{P}_{\text{rdp},2,n}$ which is a recursive dyadic partition with at most $Ck(\log n)^2$ rectangles where C is a universal constant. Equivalently, for all $\theta \in \mathbb{R}^{L_{2,n}}$ and all non negative integers r, we have

$$k^{(r)}_{\text{rdp}}(\theta) \leq C(\log n)^2 k^{(r)}_{\text{all}}(\theta). \tag{3.2}$$

We have not seen the above result (equation (3.2)) stated explicitly in the Statistics literature. It is probable that this result is known in the combinatorics or computational geometry literature. However, since we could locate an exact reference, we provide its proof in Section C.3.

REMARK 3.6. *The exponent of $\log n$, which is 1 for $d = 1$ and 2 for $d = 2$, cannot be improved in general. It is now natural to conjecture that a result like above is true for a general d where the exponent of $\log n$ is d. However, we do not know whether this is true or not. Our current proof for the $d = 2$ case breaks down and cannot be extended to higher dimensions. See Remark C.3 for more explanations on this.*

The implication of Proposition 3.9 is the following corollary for Dyadic CART.

COROLLARY 3.10. *For $d = 1$ and any integer n, there exists a universal constant $C > 0$ such that by setting $\lambda = CK_{r,1}\,\sigma^2\,\log n$ we have the following risk bound,*

$$MSE(\widehat{\theta}^{(r)}_{\text{rdp},\lambda}, \theta^*) \leq CK_{r,1}\sigma^2 \frac{k^{(r)}_{\text{all}}(\theta^*)}{N} \log \frac{n}{k^{(r)}_{\text{all}}(\theta)} \log n + \frac{C\,\sigma^2}{N}.$$

For $d = 2$ and any integer n, there exists a universal constant $C > 0$ such that by setting $\lambda = CK_{r,2}\,\sigma^2\,\log n$ we have the following risk bound,

$$MSE(\widehat{\theta}^{(r)}_{\text{rdp},\lambda}, \theta^*) \leq CK_{r,2}\,\sigma^2 \frac{k^{(r)}_{\text{all}}(\theta^*)}{N} (\log N)^3 + \frac{C\,\sigma^2}{N}.$$

To summarize, Dyadic CART attains the same rate as the ORT with an extra $\log N$ factor when $d = 1$ and with an extra $(\log N)^2$ factor when $d = 2$. We do not know whether for $d > 2$, a result for Dyadic CART analogous to Theorem 3.8 for fat partitions is possible or not.

4. Results for Multivariate Functions with Bounded Total Variation. In this section, we will describe an application of Theorem 2.1 to show that Dyadic CART of order 0 has near optimal (worst case and adaptive) risk guarantees in any dimension when we consider estimating functions with bounded total variation. Let us first define what we mean by total variation.

Let us think of $L_{d,n}$ as the d dimensional regular lattice graph. Then, thinking of $\theta \in \mathbb{R}^{L_{d,n}}$ as a function on $L_{d,n}$ we define

$$\mathrm{TV}(\theta) = \sum_{(u,v)\in E_{d,n}} |\theta_u - \theta_v| \tag{4.1}$$

where $E_{d,n}$ is the edge set of the graph $L_{d,n}$. One way to motivate the above definition is as follows. If we think $\theta[i_1, \ldots, i_n] = f(\frac{i_1}{n}, \ldots, \frac{i_d}{n})$ for a differentiable function $f : [0,1]^d \to \mathbb{R}$ then the above definition divided by n^{d-1} is precisely the Reimann approximation for $\int_{[0,1]^d} \|\nabla f\|_1$. Of course, the definition in (4.1) applies to arbitrary arrays, not just for evaluations of a differentiable function on the grid.

The usual way to estimate functions/arrays with bounded total variation is to use the Total Variation Denoising (TVD) estimator defined as follows:

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta\in\mathbb{R}^{L_{d,n}}} \big(\|y-\theta\|^2 + \lambda \mathrm{TV}(\theta)\big).$$

This estimator was first introduced in the $d = 2$ case by Rudin et al. (1992) for image denoising. This estimator is now a standard and widely succesful technique to do image denoising. In the $d = 1$ setting, it is known (see, e.g. Donoho and Johnstone (1998), Mammen and van de Geer (1997)) that a well tuned TVD estimator is minimax rate optimal on the class of all bounded variation signals $\{\theta : \mathrm{TV}(\theta) \le V\}$ for $V > 0$. It is also known (e.g, see Guntuboyina et al. (2020), Dalalyan et al. (2017), Ortelli and van de Geer (2018)) that, when properly tuned, the above estimator is capable of attaining the oracle MSE scaling like $O(\frac{k^{(0)}_{\mathrm{all}}(\theta^*)}{N})$, up to a log factor in N.

In the multivariate setting ($d \ge 2$), worst case performance of the TVD estimator has been studied in Hütter and Rigollet (2016), Sadhanala et al. (2016). These results show that like in the 1D setting, a well tuned TVD estimator is nearly (up to log factors) minimax rate optimal over the class $\{\theta \in \mathbb{R}^{L_{d,n}} : \mathrm{TV}(\theta) \le V\}$ of bounded variation signals in any dimension.

The goal of this section is to proclaim that the Dyadic CART estimator $\widehat{\theta}^{(0)}_{\mathrm{rdp},\lambda}$ enjoys similar statistical guarantees as the TVD estimator and possibly even has some advantages over TVD which are listed explicitly in Section B.3 (in the supplementary file).

4.0.1. *Adaptive Minimax Rate Optimality of Dyadic CART.* We now describe risk bounds for the Dyadic Cart estimator for bounded variation arrays. Let us define the following class

of bounded variation arrays:

$$K_{d,n}(V) = \{\theta \in L_{d,n} : \mathrm{TV}(\theta) \leq V\}$$

For any generic subset $S \subset [d]$, let us denote its cardinality by $|S|$. For any vector $x \in [n]^d$ let us define $x_S \in [n]^{|S|}$ to be the vector x restricted to the coordinates given by S. We now define

$$K^S_{d,n}(V) = \{\theta \in K_{d,n}(V) : \theta(x) = \theta(y)\, \forall x, y \ \in \ [n]^d \ \text{with} \ x_S = y_S\}$$

In words, $K^S_{n,d}(V)$ is just the set of arrays in $K_{d,n}(V)$ which are a function of the coordinates within S only. In this section, we will show that the Dyadic CART estimator is minimax rate optimal (up to log factors) over the parameter space $K^S_{d,n}(V)$ simultaneously over all subsets $S \subset [d]$. This means that the Dyadic CART performs as well as an oracle estimator which knows the subset S. This is what we mean when we say that the Dyadic CART estimator adapts to intrinsic dimensionality. To the best of our knowledge, such an oracle property in variable selection is rare in Non Parametric regression. The work in Bertin and Lecué (2008) shows a two step procedure for adapting to instrinsic dimensionality for multivariate Holder smooth function classes. The only comparable result that we are aware of for a spatially heterogenous function class is Theorem 3 in Deng and Zhang (2018) which proves a similar adaptivity result in multivariate isotonic regression.

Fix a subset $S \subset [d]$ and let $s = |S|$. Consider our Gaussian mean estimation problem where it is known that the underlying truth $\theta^* \in K^S_{d,n}(V)$. We could think of θ^* as n^{d-s} copies of a s dimensional array $\theta^*_S \in \mathbb{R}^{L_{s,n}}$. It is easy to check that $\theta^*_S \in K_{s,n}(V_S)$ where $V_s = \frac{V}{n^{d-s}}$. Estimating θ^* is equivalent to estimating the s dimensional array θ^*_S where the noise variance is now reduced to $\sigma^2_S = \frac{\sigma^2}{n^{d-s}}$ because we can average over n^{d-s} elements per each entry of θ^*_S. Therefore, we now have a reduced Gaussian mean estimation problem where the noise variance is σ^2_S and the parameter space is $K_{n,s}(V_S)$. A tight lower bound to the minimax risk for the parameter space $K_{d,n}(V)$ for arbitrary $n, d, V > 0$ is available in Sadhanala et al. (2016). Using the above logic and this existing minimax lower bound allows us to establish a lower bound to the minimax risk for the parameter space $K^S_{d,n}(V)$. The detailed proof is given in Section C.

THEOREM 4.1 (Minimax Lower Bound over $K^S_{d,n}(V)$). *Fix positive integers n, d and let $S \subset [d]$ such that $s = |S| \geq 2$. Let $V > 0$ and $V_S = \frac{V}{n^{d-s}}$. Similarly, for $\sigma > 0$, let $\sigma^2_S = \frac{\sigma^2}{n^{d-s}}$. There exists a universal constant $c > 0$ such that*

$$\inf_{\widetilde{\theta} \in \mathbb{R}^{L_{d,n}}} \sup_{\theta \in K^S_{d,n}(V)} \mathbb{E}_\theta \|\widetilde{\theta} - \theta\|^2 \geq c\, n^{d-s}\ \min\{\frac{\sigma_S\, V_S}{2s}\sqrt{1 + \log(\frac{2\,\sigma\, s\, n^s}{V_S})}, n^s \sigma^2_S, \frac{{V_S}^2}{s} + \sigma^2_S\}.$$

If $|S| = 1$ then

$$\inf_{\widetilde{\theta} \in \mathbb{R}^{L_{d,n}}} \sup_{\theta \in K^S_{d,n}(V)} \mathbb{E}_\theta \|\widetilde{\theta} - \theta\|^2 \geq c\, n^{d-1}\ \min\{(\sigma^2_S V_S)^{2/3} n^{1/3}, n\, \sigma^2_S, n\, V^2_S\}.$$

Let us now explain the above result. If we take the subset $S = [d]$ this is exactly the lower bound in Theorem 2 of Sadhanala et al. (2016). All we have done is stated the same result for any subset S since we can reduce the estimation problem in $K^S_{d,n}(V)$ to a s dimensional estimation problem over $K_{s,n}(V_S)$. The bound is in terms of a minimum of three terms. It is enough to explain this bound in the case when $S = [d]$ as similar reasoning holds for any subset S with $s = |S| \geq 2$. Thinking of σ as a fixed constant, the three terms in the minimum on the right side corresponds to different regimes of V. It can be shown that the constant array with each entry $\overline{y}$ attains the $V^2 + \sigma^2$ rate which is dominant when V is very small. The estimator y itself attains the $N\sigma^2$ rate which is dominant when V is very large. Hence, these regimes of V can be thought of as trivial regimes. In the nontrivial regime, the lower bound is $c \min\{\frac{\sigma V}{2d}\sqrt{1 + \log(\frac{2\,\sigma\, d\, N}{V})}\}$.

It is also known that a well tuned TVD estimator is minimax rate optimal, in the nontrivial regime, over $K_{d,n}(V)$ for all $d \geq 2$, up to log factors; see Hütter and Rigollet (2016). For instance, it achieves the above minimax lower bound (up to log factors) in the nontrivial regime. For this reason, we can define an oracle estimator (which knows the set S) attaining the minimax lower bound over $K^S_{d,n}(V)$ in Theorem 4.1, up to log factors. The oracle estimator would first obtain $\overline{y}_S$ by averaging the observation array y over the coordinates in S^C and then it would apply the s dimensional TVD estimator on $\overline{y}_S$. Our main point here is that the Dyadic CART estimator performs as well as this oracle estimator, without the knowledge of S. In other words, its risk nearly (up to log factors) matches the minimax lower bound in Theorem 4.1 adaptively over all subsets $S \subset [d]$. This is the content of our next theorem which is proved in Section C (in the supplementary file).

THEOREM 4.2 (Adaptive Risk Bound for Dyadic Cart). *Fix any positive integers n, d. Let $\theta^* \in K^S_{d,n}(\infty)$ be the underlying truth where $S \subset [d]$ is any subset with $|S| \geq 2$. Let $V^* = \mathrm{TV}(\theta^*)$. Let $V^*_S = \frac{V^*}{n^{d-s}}$ and $\sigma^2_S = \frac{\sigma^2}{n^{d-s}}$ be defined as before. The following risk bound holds for the Dyadic CART estimator $\widehat{\theta}^{(0)}_{\mathrm{rdp},\lambda}$ with $\lambda \geq C\sigma^2 \log N$ where C is an absolute constant.*

$$\mathbb{E}_{\theta^*}\|\widehat{\theta}^{(0)}_{\mathrm{rdp},\lambda} - \theta^*\|^2 \leq C\, n^{d-s}\, \min\{\sigma_S V^*_S \log N, \sigma^2_S \log N, \big((V^*_S)^2 + \sigma^2_S\big)\}$$

In the case $|S| = 1$ we have

$$\mathbb{E}_{\theta^*}\|\widehat{\theta}^{(0)}_{\mathrm{rdp},\lambda} - \theta^*\|^2 \leq C\, n^{d-1}\, \min\{(\sigma^2_S V_S \log N)^{2/3} n^{1/3}, n\,\sigma^2_S \log N, n\, V^2_S + \sigma^2_S \log N\}$$

We think the following is an instructive way to read off the implications of the above theorem. Let us consider $d \geq 2$ and the $S = [d]$ case. We will only look at the nontrivial regime even though Dyadic CART remains minimax rate optimal, up to log factors, even in the trivial regimes. In this case, $MSE(\widehat{\theta}^{(0)}_{\mathrm{rdp},\lambda}, \theta^*) = \widetilde{O}(\frac{\sigma V^*}{N})$ which is the minimax rate in the nontrivial regime as given by Theorem 4.1. Now, for many natural instances of θ^*, the quantity $V^* = O(n^{d-1})$; for instance if θ^* are evaluations of a differentiable function on the grid. This $O(n^{d-1})$ scaling was termed as the *canonical scaling* for this problem

by Sadhanala et al. (2016). Therefore, under this canonical scaling for V^* we have

$$MSE(\widehat{\theta}^{(0)}_{\text{rdp},\lambda}, \theta^*) = \widetilde{O}(\frac{\sigma}{n}) = \widetilde{O}(\frac{\sigma}{N^{1/d}}).$$

Now let us consider $d \geq 2$ and a general subset $S \subset [d]$. In the nontrivial regime, by Theorem 4.2 we have $MSE(\widehat{\theta}^{(0)}_{\text{rdp},\lambda}, \theta^*) = \widetilde{O}(\frac{\sigma_S V_S^*}{n^s})$ which is also the minimax rate over the parameter space $K^S_{d,n}$. Now, $V_S^* = O(n^{s-1})$ under the canonical scaling in this case. Thus, under this canonical scaling we can write

$$MSE(\widehat{\theta}^{(0)}_{\text{rdp},\lambda}, \theta^*) = \widetilde{O}(\frac{\sigma_S}{n}) = \widetilde{O}(\frac{\sigma_S}{N^{1/d}}).$$

This is very similar to the last display except σ has been replaced by σ_S, the *actual standard deviation* of this problem. The point is, the Dyadic CART attains this rate without knowing S. The case when $|S| = 1$ can be read off in a similar way.

5. Results for Univariate Functions of Bounded Variation of Higher Orders. In this section, we show another application of Theorem 2.1 to a family of univariate function classes which have been of recent interest. The results in this section would be for the univariate Dyadic Cart estimator of some order $r \geq 0$. As mentioned in Section 1, TV denoising in the 1D setting has been studied as part of a general family of estimators which penalize discrete derivatives of different orders. These estimators have been studied in Mammen and van de Geer (1997) , Steidl et al. (2006), Tibshirani (2014), Guntuboyina et al. (2020) and Kim et al. (2009) who coined the name trend filtering.

To define the trend filtering estimators here, we first need to define variation of all orders. For a vector $\theta \in \mathbb{R}^n$, let us define $D^{(0)}(\theta) = \theta, D^{(1)}(\theta) = (\theta_2 - \theta_1, \ldots, \theta_n - \theta_{n-1})$ and $D^{(r)}(\theta)$, for $r \geq 2$, is recursively defined as $D^{(r)}(\theta) = D^{(1)}(D^{(r-1)}(\theta))$. Note that $D^{(r)}(\theta) \in \mathbb{R}^{n-r}$. For simplicity, we denote the operator $D^{(1)}$ by D. For any positive integer $r \geq 1$, let us also define the r th order variation of a vector θ as follows:

$$V^{(r)}(\theta) = n^{r-1}|D^{(r)}(\theta)|_1 \tag{5.1}$$

where $|.|_1$ denotes the usual ℓ_1 norm of a vector. Note that $V^{(1)}(\theta)$ is the usual total variation of a vector as defined in (4.1).

REMARK 5.1. *The n^{r-1} term in the above definition is a normalizing factor and is written following the convention adopted in Guntuboyina et al. (2020). If we think of θ as evaluations of a r times differentiable function $f : [0,1] \to \mathbb{R}$ on the grid $(1/n, 2/n \ldots, n/n)$ then the Reimann approximation to the integral $\int_{[0,1]} f^{(r)}(t)dt$ is precisely equal to $V^{(r)}(\theta)$. Here $f^{(r)}$ denotes the rth derivative of f. Thus, for natural instances of θ, the reader can imagine that $V^{(r)} = O(1)$.*

Let us now define the following class of sequences for any integer $r \geq 1$,

$$\mathcal{BV}^{(r)}_n(V) = \{\theta \in \mathbb{R}^n : V^{(r)}(\theta) \leq V\}. \tag{5.2}$$

Trend Filtering (of order $r \geq 1$) estimators are defined as follows for a tuning parameter $\lambda > 0$:

$$\widehat{\theta}^{(r)}_{tf,\lambda} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \left(\|y - \theta\|^2 + \lambda V^{(r)}(\theta) \right).$$

Thus, Trend Filtering is penalized least squares where the penalty is proportional to the ℓ_1 norm of $D^{(r)}(\theta)$. As opposed to Trend Filtering, here we will study the univariate Dyadic CART estimator (of order $r-1$) which penalizes something similar to the ℓ_0 norm of $D^{(r)}(\theta)$. The results presented in this section will show that the Dyadic CART (of order $r-1$) compares favourably with Trend Filtering (of order r) in several aspects as listed in Section B.4 (in the supplementary file).

5.0.1. *Risk Bounds for Univariate Dyadic CART of all orders.* We start with the bound of $n^{-2r/(2r+1)}$ for the risk of Dyadic CART of order $r-1$ for the parameter space $\mathcal{BV}^{(r)}_n(V)$. We also explicitly state the dependence of the bound on V and σ.

THEOREM 5.1 (Slow Rate for Dyadic CART). *Fix a positive integer r. Let $V^r(\theta^*) = V$. For the same constant C as in Theorem 2.1, if we set $\lambda \geq C\sigma^2 \log n$ we have*

$$MSE(\widehat{\theta}^{(r-1)}_{\text{rdp},\lambda}, \theta^*) \leq C_r \Big(\frac{\sigma^2 V^{1/r} \log n}{n} \Big)^{2r/(2r+1)} + C_r \sigma^2 \frac{\log n}{n} \tag{5.3}$$

where C_r is an absolute constant only depending on r.

REMARK 5.2. *The proof of the above theorem is done in Section C (in the supplementary file). The proof proceeds by approximating any $\theta \in \mathcal{BV}^{(r)}_n(V)$ with a vector θ' which is piecewise polynomial of degree $r-1$ with an appropriate bound on its number of pieces and then invoking Theorem 2.1.*

REMARK 5.3. *The above theorem shows that the univariate Dyadic CART estimator of order $r-1$ is minimax rate optimal up to the $(\log n)^{2r/(2r+1)}$ factor. The dependence of V is also optimal in the above bound. Up to the log factor, this upper bound matches the bound already known for the Trend Filtering estimator of order r; (see e.g, Tibshirani (2014)).*

Our next bound shows that the univariate Dyadic CART estimator achieves our goal of attaining the oracle risk for piecewise polynomial signals.

THEOREM 5.2 (Fast Rates for Dyadic CART). *Fix a positive integer r and $0 < \delta < 1$. Let $V^r(\theta^*) = V$. For the same constant C as in Theorem 2.1, if we set $\lambda \geq C\sigma^2 \log n$ we have*

$$\mathbb{E}\|\widehat{\theta}^{(r)}_{\text{rdp},\lambda} - \theta^*\|^2 \leq \inf_{\theta \in \mathbb{R}^N} \Big[\frac{(1+\delta)}{(1-\delta)} \|\theta - \theta^*\|^2 + \frac{\lambda C_r}{1-\delta} k^{(r)}_{\text{all}}(\theta) \log\Big(\frac{en}{k^{(r)}_{\text{all}}(\theta)}\Big) \Big] + C \frac{\sigma^2}{\delta(1-\delta)}$$

where C_r is an absolute constant only depending on r. As a corollary we can conclude that

$$MSE(\widehat{\theta}^{(r)}_{\text{rdp},\lambda}, \theta^*) \leq C_r \sigma^2 \frac{k^{(r)}_{\text{all}}(\theta^*) \log n \log\big(\frac{en}{k^{(r)}_{\text{all}}(\theta^*)}\big)}{n}.$$

PROOF. The proof follows directly from the risk bound for univariate Dyadic Cart given in Theorem 2.1 and applying equation (3.1) in Lemma C.3 which says that $k^{(r)}_{\text{rdp}}(\theta) \leq k^{(r)}_{\text{all}}(\theta) \log(\frac{en}{k^{(r)}_{\text{all}}(\theta)})$ for all vectors $\theta \in \mathbb{R}^n$. □

Let us now put our result in Theorem 5.2 in context. It says that in the $d = 1$ case, Dyadic CART achieves our goal of attaining MSE scaling like $\widetilde{O}(k^{(r)}_{\text{all}}(\theta^*)/n)$ (fast rate) for all θ^*. The Trend Filtering estimator, ideally tuned, is also capable of attaining this rate of convergence; (see Theorem 3.1 in van de Geer and Ortelli (2019) and Theorem 2.1 in Guntuboyina et al. (2020)), under certain *minimum length conditions* on θ^*. However, Dyadic CART does not need such minimum length conditions for the fast rate to hold. This issue is discussed in more detail in Section B.5 (in the supplementary material) which includes comparisons between Theorem 5.2 and the comparable result known for Trend Filtering.

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to "Adaptive Estimation of Multivariate Piecewise Polynomials and Bounded Variation Functions by Optimal Decision Trees"
(). This supplementary material contains simulations, discussion about our results and the proofs of our theorems and auxiliary results.

References.

Berman, P., B. DasGupta, and S. Muthukrishnan (2002). Exact size of binary space partitionings and improved rectangle tiling algorithms. *SIAM Journal on Discrete Mathematics 15*(2), 252–267.

Bertin, K. and G. Lecué (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics 2*, 1224–1241.

Bertsimas, D. and J. Dunn (2017). Optimal classification trees. *Machine Learning 106*(7), 1039–1082.

Blanchard, G., C. Schäfer, Y. Rozenholc, and K.-R. Müller (2007). Optimal dyadic decision trees. *Machine Learning 66*(2-3), 209–241.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). Classification and regression trees. wadsworth int. *Group 37*(15), 237–251.

Chatterjee, S. and S. Goswami (2019). New risk bounds for 2d total variation denoising. *arXiv preprint arXiv:1902.01215*.

Chaudhuri, P., M.-C. Huang, W.-Y. Loh, and R. Yao (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, 143–167.

Dalalyan, A., M. Hebiri, and J. Lederer (2017). On the prediction performance of the lasso. *Bernoulli 23*(1), 552–581.

de Berg, M. (1995). Linear size binary space partitions for fat objects. In *European Symposium on Algorithms*, pp. 252–263. Springer.

Deng, H. and C.-H. Zhang (2018). Isotonic regression in multi-dimensional spaces and graphs. *arXiv preprint arXiv:1812.08944*.

Donoho, D. L. (1997). CART and best-ortho-basis: a connection. *The Annals of Statistics 25*(5), 1870–1911.

Donoho, D. L. and I. M. Johnstone (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics 26*(3), 879–921.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Gey, S. and E. Nedelec (2005). Model selection for cart regression trees. *IEEE Transactions on Information Theory 51*(2), 658–670.

Guntuboyina, A., D. Lieu, S. Chatterjee, and B. Sen (2020). Adaptive risk bounds in univariate total variation denoising and trend filtering. *The Annals of Statistics 48*(1), 205–229.

Hershberger, J., S. Suri, and C. D. Tóth (2005). Binary space partitions of orthogonal subdivisions. *SIAM Journal on Computing 34*(6), 1380–1397.

Hütter, J.-C. and P. Rigollet (2016). Optimal rates for total variation denoising. In *Conference on Learning Theory*, pp. 1115–1146.

Ishwaran, H. (2015). The effect of splitting on random forests. *Machine Learning 99*(1), 75–118.

Kim, S.-J., K. Koh, S. Boyd, and D. Gorinevsky (2009). ℓ_1 trend filtering. *SIAM Rev. 51*(2), 339–360.

Laurent, H. and R. L. Rivest (1976). Constructing optimal binary decision trees is np-complete. *Information processing letters 5*(1), 15–17.

Lecué, G. (2008). Classification with minimax fast rates for classes of bayes rules with sparse representation. *Electronic Journal of Statistics 2*, 741–773.

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review 82*(3), 329–348.

Mammen, E. and S. van de Geer (1997). Locally adaptive regression splines. *The Annals of Statistics 25*(1), 387–413.

Nemirovski, A. (2000). Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour 28*, 85.

Nowak, R., U. Mitra, and R. Willett (2004). Estimating inhomogeneous fields using wireless sensor networks. *IEEE Journal on Selected Areas in Communications 22*(6), 999–1006.

Ortelli, F. and S. van de Geer (2018). On the total variation regularized estimator over a class of tree graphs. *Electron. J. Statist. 12*(2), 4517–4570.

Rudin, L. I., S. Osher, and E. Fatemi (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena 60*(1), 259–268.

Sadhanala, V., Y.-X. Wang, and R. J. Tibshirani (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, pp. 3513–3521.

Scornet, E., G. Biau, and J.-P. Vert (2015). Consistency of random forests. *The Annals of Statistics 43*(4), 1716–1741.

Scott, C. and R. D. Nowak (2006). Minimax-optimal classification with dyadic decision trees. *IEEE transactions on information theory 52*(4), 1335–1353.

Steidl, G., S. Didas, and J. Neumann (2006). Splines in higher order tv regularization. *International journal of computer vision 70*(3), 241–255.

Tibshirani, R. (2015). Nonparametric regression (and classification).

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics 42*(1), 285–323.

Tóth, C. D. (2005). Binary space partitions: recent developments. *Combinatorial and Computational Geometry. MSRI Publications 52*, 529–556.

van de Geer, S. and F. Ortelli (2019). Prediction bounds for (higher order) total variation regularized least squares. *arXiv preprint arXiv:1904.10871*.

Wager, S. and G. Walther (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Willett, R. M. and R. D. Nowak (2007). Multiscale poisson intensity and density estimation. *IEEE Transactions on Information Theory 53*(9), 3171–3187.