Electronic Journal of Statistics
Vol. 18 (2024) 1206–1247
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2214

Quantile regression by dyadic CART

Oscar Hernan Madrid Padilla

Department of Statistics
University of California, Los Angeles
520 Portola Plaza, Los Angeles, California 90095, U.S.A.
e-mail: oscar.madrid@stat.ucla.edu

and

Sabyasachi Chatterjee

Department of Statistics
University of Illinois at Urbana-Champaign
725 S. Wright St. M/C 374, Champaign, Illinois 61820, U.S.A.
e-mail: sc1706@illinois.edu

Abstract: In this paper we propose and study a version of the Dyadic Classification and Regression Trees (DCART) estimator from Donoho (1997) for (fixed design) quantile regression in general dimensions. We refer to this proposed estimator as the QDCART estimator. Just like the mean regression version, we show that a) a fast dynamic programming based algorithm with computational complexity $O(N \log N)$ exists for computing the QDCART estimator and b) an oracle risk bound (trading off squared error and a complexity parameter of the true signal) holds for the QDCART estimator. This oracle risk bound then allows us to demonstrate that the QDCART estimator enjoys adaptively rate optimal estimation guarantees for piecewise constant and bounded variation function classes. In contrast to existing results for the DCART estimator which requires subgaussianity of the error distribution, for our estimation guarantees to hold we do not need any restrictive tail decay assumptions on the error distribution. For instance, our results hold even when the error distribution has no first moment such as the Cauchy distribution. Furthermore, we perform extensive numerical experiments on both simulated and real data which illustrate the usefulness of the proposed methods.

MSC2020 subject classifications: Primary 62G08; secondary 62G05.
Keywords and phrases: Classification and regression trees (CART), recursive dyadic partitions, piecewise constant signals, dynamic programming.

Received June 2022.

Contents

1. Introduction

We consider the problem of nonparametric quantile regression in general dimensions and specifically consider the setting of fixed/lattice design regression or array denoising. In this setting, we are given an array of independent random variables $y \in \mathbb{R}^{L_{d,n}}$ where $L_{d,n}$ is the d-dimensional square lattice or grid graph with nodes indexed by $\{1, \ldots, n\}^d$. Then the goal is to estimate the *true* τ *quantile array* θ^* where

$$\theta_i^* = \arg\min_{a \in \mathbb{R}} \mathbb{E}\left(\rho_\tau\left(y_i - a\right)\right)$$

for all $i \in L_{d,n}$, $\tau \in (0,1)$ is a fixed quantile level and where $\rho_\tau(x) = \max\{\tau x, (1-\tau)x\}$ is the usual piecewise linear convex quantile loss function. For example, when $\tau = 0.5$, our setting here amounts to estimate the true median array of the noisy array y. The model here can be called *the quantile sequence model*. This generalizes the usual Gaussian sequence model where the quantile τ is taken to be 0.5 and the distribution of y is taken to be multivariate normal with the covariance matrix a multiple of identity.

Assuming lattice design is common practice for studying non parametric regression estimators and clearly our setting is relevant for image denoising and computer vision when $d = 2$ or 3. The problem of estimating the true signal θ^* becomes meaningful when the true array satisfies some additional structure so that the effective parameter size is much less even though the actual number of unknown parameters is the same as the sample size. Structured signal denoising

is a standard problem and arises in several scientific disciplines, e.g see applications in computer vision (e.g. Bian et al., 2017; Wirges et al., 2018), medical imaging (e.g. Lang et al., 2014), and neuroscience (e.g. Tansey et al., 2018).

In this paper we are interested in scenarios where θ^* is (or is close to) a piecewise constant array on a rectangular partition of $L_{d,n}$. For mean regression, Dyadic classification and regression trees (DCART) method introduced in Donoho (1997) is known to be computationally efficient while achieving adaptively minimax rate optimal rates of convergence for classes of signals which are piecewise constant in a rectangular partition of $L_{d,n}$; see Chatterjee and Goswami (2019) for a thorough study of statistical adaptivity of DCART. However, since we are interested in quantile regression, we would like to propose a quantile version of Dyadic CART.

The most natural way to define the quantile version of Dyadic CART estimator is as follows:

$$\hat{\theta}_{rdp} = \underset{\theta \in \mathbb{R}^{L_{d,n}}}{\arg\min} \left\{ \sum_{i \in L_{d,n}} \rho_\tau(y_i - \theta_i) + \lambda k_{rdp}(\theta) \right\}, \tag{1.1}$$

where we define $k_{rdp}(\theta)$ as the smallest natural number for which there exists a dyadic partition Π of $L_{d,n}$ such that θ is constant in each element of Π and $|\Pi| = k_{rdp}(\theta)$. The estimator we propose and study in this article is a slightly modified version of the above estimator in (1.1). We refer to this proposed estimator as the QDCART estimator. The precise definition of our estimator and the meaning of a dyadic rectangular partition of $L_{d,n}$ and the complexity parameter $k_{rdp}(\theta)$ will be given in Section 2.

The usual mean regression version of Dyadic CART estimator is a computationally feasible decision tree method proposed first in Donoho (1997) in the context of regression on a two-dimensional grid design. This estimator optimizes the same criterion as in (1.1) except that the quantile loss is replaced by the usual squared loss. Subsequently after Donoho (1997), several papers have used ideas related to dyadic partitioning for regression, classification and density estimation; e.g see Nowak et al. (2004), Scott and Nowak (2006), Blanchard et al. (2007), Willett and Nowak (2007). Recently, the paper Chatterjee and Goswami (2019) generalized the Dyadic CART estimator to general dimensions and to higher orders and studied the ability of Dyadic CART to estimate piecewise constant signals of various types. Dyadic CART has also been recently used for recovering level sets of piecewise constant signals; see Padilla et al. (2021). It is fair to say that the two most important facts about the usual mean regression version of Dyadic CART are:

- The Dyadic CART estimator attains an oracle risk bound; e.g see Theorem 2.1 in Chatterjee and Goswami (2019). This oracle risk bound can then be used to show that the Dyadic CART estimator is nearly minimax rate optimal for several function classes of interest.
- The Dyadic CART estimator can be computed by a bottom up dynamic program with computational complexity linear in the sample size, see

Lemma 1.1 in Chatterjee and Goswami (2019).

These two properties of the Dyadic CART make it a very attractive signal denoising method. However, the oracle risk bound satisfied by Dyadic CART is known to hold only under sub-Gaussian errors. A natural question is whether it is possible to define a version of Dyadic CART which satisfies a result like Theorem 2.1 in Chatterjee and Goswami (2019) without any tail decay assumptions on the error distribution and still retains essentially linear time computational complexity? This is the main question that motivated the research in this article and naturally led us to study a quantile regression version of Dyadic CART. The results in this paper answer our question as affirmative. We now summarize our results.

- Theorem 1 gives an oracle risk bound for the QDCART estimator proposed in this paper. The advantage of our risk bound is that it holds under an extremely mild assumption (see Assumption 1 in Section 3) on the distribution of the error or noise variables. For example, our risk bound holds when the error distribution is heavy tailed like the Cauchy distribution for which even the first moment does not exist. In contrast, Theorem 2.1 in Chatterjee and Goswami (2019) heavily relies on the subgaussian nature of the errors. Therefore, our main contribution here is to establish the robustness of the quantile version of Dyadic CART to heavy tailed errors. The result in Theorem 1 can be thought of as generalizing Theorem 2.1 in Chatterjee and Goswami (2019) to the heavy tailed setting. First, we show in Lemma 1 that the estimator has bounded ℓ_∞ norm with high probability. Then, in Lemma 2, we show a modified version of Lemma 9.1 in Chatterjee and Goswami (2019), the first key empirical process control in our analysis. Next we use symmetrization (Lemma 3) and the contraction principle (Lemma 4) to control the second key empirical process in our analysis (see Lemma 7). The latter is combined with a careful peeling argument in the proof of Theorem 4 exploiting the properties of the quantile loss as stated in Lemmas 5 and 8. Thus at a high level, our proof has several extra steps in comparison to the analysis in Chatterjee and Goswami (2019) as the authors there mainly had to focus on upper bounding a empirical process as in Lemma 9.1 therein. These extra steps are necessary precisely because we are allowing arbitrarily heavy tailed errors. At a high level, what makes it possible to handle arbitrarily heavy tailed errors are that we are doing penalized quantile loss regression and not penalized square loss regression, the key empirical process comes out in terms of Rademacher random variables irrespective of the original error distribution, and sample quantiles concentrate (as long as the sample size is not too few) for arbitrarily heavy tailed distributions.
- Once the oracle risk bound in Theorem 1 has been established, it has been shown in Chatterjee and Goswami (2019) how this automatically implies that the QDCART estimator would be minimax rate optimal for several function/signal classes of interest. In particular, this opens the door for us to establish minimax rate optimality of our QDCART estimator

over the space of piecewise constant and/or bounded variation arrays. We provide these results in Section 3.1. At the risk of reiterating, the state of the art mean regression estimators for estimating piecewise constant and/or bounded variation arrays typically require subgaussianity of the errors while the QDCART estimator is robust to heavy tailed error distributions. A natural competing quantile regression estimator to QDCART is the Quantile Total Variation Denoising estimator studied in Padilla and Chatterjee (2020). Just like for the corresponding mean regression counterparts, we argue in Section 3.1 that the QDCART estimator has certain advantages over the Quantile Total Variation Denoising estimator, not least the fact that QDCART is computable in essentially linear time in any dimension whereas Quantile Total Variation Denoising is not known to have linear time computational complexity in multivariate settings ($d > 1$).

- We explain in Section A that our proof technique for Theorem 1 can also be used to derive similar risk bounds for other variants of the QDCART estimator. For example, in Chatterjee and Goswami (2019) the Optimal Regression Tree (ORT) estimator was introduced and studied for mean regression. This ORT estimator is similar to the Dyadic CART estimator with the same optimization objective function except that the optimization is done over all decision trees or hierarchical partitions (not necessarily dyadic). It was then shown in Chatterjee and Goswami (2019) that this estimator attains a better risk bound than Dyadic CART in general. However, its computational complexity is slower and scales like $O(N^{2+1/d})$ in d dimensions in contrast to the $O(N)$ computational complexity of Dyadic CART. The proof techniques of this paper actually also imply that a quantile version of the ORT estimator can be defined which will enjoy the corresponding risk guarantee. We prefer to present our main results only for QDCART to make the exposition short and because of its significantly better computational complexity.
- We give a bottom up dynamic programming algorithm which can exactly compute the QDCART estimator. This algorithm is similar to the original one proposed for the DCART estimator in Donoho (1997), suitably adapted to our setting. The computational complexity of our algorithm is $O(N(\log N)^d)$ (see Theorem 2) which is slightly slower than the $O(N)$ computational complexity of the DCART estimator. This extra log factor in the computation seems unavoidable to us because of the need to compute and propagate quantiles of various dyadic rectangles. Our algorithm is described in detail in Section 5.

1.1. Outline

The rest of the paper is organized as follows. Section 2 presents the precise definition of the QDCART estimator. The main theoretical result (Theorem 1) of this paper is then presented in Section 3. We then provide implications of our

main result (Theorem 1) to the class of bounded variation signals in Section 3.1, and to the class of piecewise constant signals in Section 3.2. Section 4 is a discussion section. Section 5 provides the details of our algorithm for implementing QDCART. Section 6 contains extensive numerical results in both simulated and real data examples.

2. Description of QDCART estimator

In this section, we precisely describe the QDCART estimator we propose to study. Let's first introduce some notation which we will use throughout this article. For any fixed dimension $d \geq 1$, we denote our *sample size* by $N = n^d$ which is the size of the lattice $L_{d,n}$. Let us denote the discrete interval of positive integers as $[a, b] := \{i \in \mathbb{Z}_+ : a \leq i \leq b\}$ where $\mathbb{Z}_+$ denotes the set of positive integers. For a positive integer n we also denote the set $[1, n]$ by just $[n]$. For squences a_n and b_n we write $a_n = O(b_n)$ if there exists a positive constant $c > 0$ such that $a_n \leq cb_n$. If instead $a_n \leq b_n(\log n)^l$ for a positive constant l then we write $a_n = \tilde{O}(b_n)$. A subset $R \subset L_{d,n}$ is called an *axis aligned rectangle* if R is a product of discrete 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. The size of a rectangle $R = \prod_{i=1}^{d}[a_i, b_i]$ is denoted by $|R|$ and defined as

$$|R| = \prod_{i=1}^{d}(b_i - a_i + 1).$$

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) $\cup_{R\in\mathcal{R}}R = L_{d,n}$.

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 intervals fo equal size. We assume that the interval has even size for ease of exposition. If not, then one can set forth a convention for defining the middle point and then follow it throughout. A dyadic partition of $L_{d,n}$ is constructed iteratively as follows. 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 $\mathcal{P}_{rdp}(L_{d,n})$. Figure 1 shows a depiction of a dyadic partition.

For a given array $\theta \in \mathbb{R}^{L_{d,n}}$, let $k_{rdp}(\theta)$ *denote the smallest positive integer* k *such that a set of* k *rectangles* $R_1, \ldots, R_k$ *form a recursive dyadic partition of* $L_{d,n}$ *and the restricted array* θ_{R_i} *is a constant array for all* $1 \leq i \leq k$. In other words, $k_{rdp}(\theta)$ is the cardinality of the minimal recursive dyadic partition of $L_{d,n}$ such that θ is piecewise constant on the partition. A visual representation of $k_{rdp}(\theta)$ is given in Figure 2 for a signal $\theta \in \mathbb{R}^{L_{2,n}}$.

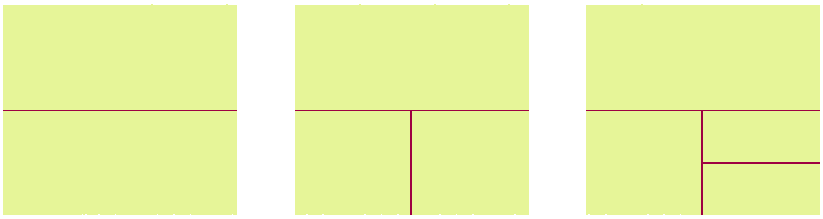

FIG 1. *From left to right the panels show an example of a sequence of three dyadic splits that lead to a dyadic parition.*

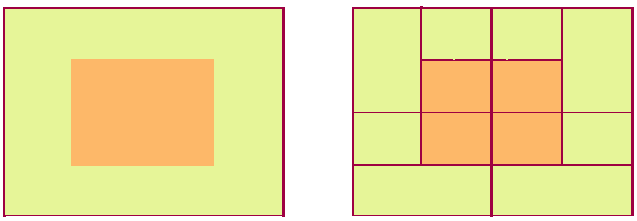

FIG 2. *The left panel shows the representation of a* $\theta \in \mathbb{R}^{L_{2,n}}$ *that takes on two values. The right panel shows a dyadic partition with a minimal number of elements where* θ *is piecewise constant. In this example* $k_{rdp}(\theta) = 12$*, the number of rectangles in the dyadic partition in the right panel.*

To define our estimator, we will need a few more notations. If Π is any rectangular partition of $L_{d,n}$ we let $S(\Pi)$ be the linear subspace of $\mathbb{R}^{L_{d,n}}$ consisting of vectors with constant values on each rectangle of Π. We also write $R \in \Pi$ to mean that the rectangle R is one of the constituent rectangles of the partition Π. We now define $O_{\Pi,\tau}(\cdot)$ be a function from $\mathbb{R}^{L_{d,n}}$ to $S(\Pi)$ such that

$$(O_{\Pi,\tau}(y))_i = q_\tau(y_R)$$

for $i \in R$, $R \in \Pi$, and where $q_\tau(y_R)$ is the empirical τ-quantile of the set of values $y_R := (y_i)_{i \in R}$.

Armed with the above notation we can reformulate the optimization problem in (1.1) by noting that $\hat{\theta}$ defined in (1.1) is the same as $O_{\tilde{\Pi},\tau}(y)$ where the partition $\tilde{\Pi}$ is an optimal solution to the following discrete optimization problem:

$$\min_{\Pi \in \mathcal{P}_{rdp}(L_{d,n})} \left\{ \sum_{i \in L_{d,n}} \rho_\tau(y_i - (O_{\Pi,\tau}(y))_i) + \lambda |\Pi| \right\}. \tag{2.1}$$

However, the estimator defined in (1.1) is not quite the estimator we propose and study in this paper as we need to modify the estimator slightly. To describe our QDCART estimator, which is the main object of study in this paper, we now define for any fixed quantile level $0 < \tau < 1$,

$$\hat{\theta}_{rdp} = O_{\hat{\Pi},\tau}(y) \tag{2.2}$$

where

$$\hat{\Pi} := \underset{\Pi \in \mathcal{P}_{rdp}(L_{d,n}) : |R| \geq \gamma \ \forall R \in \Pi}{\arg\min} \left\{ \sum_{i \in L_{d,n}} \rho_\tau(y_i - (O_{\Pi,\tau}(y))_i) + \lambda |\Pi| \right\} \tag{2.3}$$

for tuning parameters $\lambda, \gamma > 0$.

Note that in view of (2.1), the above QDCART estimator is basically the same as the estimator in (1.1) with a slight modification. We restrict the optimization space to all partitions in $\mathcal{P}_{rdp}(L_{d,n})$ with the constraint that the size of each of its constituent rectangles is larger than $\gamma > 0$. This restriction is needed to avoid the estimator from being affected by large outliers. We say more on this point in Remark 1.

3. Main results

We first state an assumption on the distribution of the coordinates of the data vector y.

Assumption 1. *There exist positive constants L, $\underline{f}$ and $\overline{f}$ such that for any $\delta \in \mathbb{R}^{L_{d,n}}$ satisfying $\|\delta\|_\infty \leq L$ we have that for $i \in L_{d,n}$,*

$$\overline{f}\,|\delta_i| \geq |F_{y_i}(\theta_i^* + \delta_i) - F_{y_i}(\theta_i^*)| \geq \underline{f}\,|\delta_i|, \tag{3.1}$$

where F_{y_i} is the cumulative distribution function of y_i.

Assumptions like the above are standard and commonly made in the quantile regression literature. For instance, without the upper bound part, Assumption 1 (Equation (3.1)) also appeared in Padilla and Chatterjee (2020) and is a weaker version of Condition 2 from He and Shi (1994), and is closely related to Condition D.1 in Belloni and Chernozhukov (2011). The lower bound in Assumption 1 is needed to ensure uniqueness of the τ quantiles of the marginal distributions of the coordinates of y. We believe that Assumption 1 is very mild. For example, sequences of distributions which are stochastically dominated by a distribution with continuous density (w.r.t. Lebesgue measure) which is bounded away from 0 on any compact interval satisfy Assumption 1. If it is assumed that the errors are i.i.d. (which is a commonly made assumption) then if the error distribution itself has continous density (w.r.t Lebesgue measure) which is bounded away from 0 on any compact interval then Assumption 1 is satisfied. In particular, the error distributions could be i.i.d. Cauchy with no moments existing.

Before stating our main result we will need to make the following definition.

Definition 1. *Let $b > 1$ be fixed and θ' and θ'' be arrays of the true τ/b-quantiles and $(1-\tau)/b$-quantiles of y respectively, so that*

$$\theta'_i = \underset{a \in \mathbb{R}}{\arg\min}\, \mathbb{E}\,\{\rho_{\tau/b}(y_i - a)\}, \quad \text{and} \quad \theta''_i = \underset{a \in \mathbb{R}}{\arg\min}\, \mathbb{E}\,\{\rho_{(1-\tau)/b}(y_i - a)\}.$$

Then we denote $U := \max\{\|\theta'\|_\infty, \|\theta''\|_\infty\}$.

Definition 1 simply quantifies the supremum norm of the τ/b-quantiles and $(1-\tau)/b$-quantiles of y. We are now ready to state our main result for the QDCART estimator.

Theorem 1. *Suppose that Assumption 1 holds. There exists universal constants $c_1, C_1, C_2, C_3 > 0$ such that for any $0 < \epsilon < 1$, if we set $\gamma = c_1 \log N$ and $\lambda = C_1 \max\{1, U\} \log(N) \log(NU)/\epsilon$, then with probability at least $1 - C_2\epsilon$,*

$$\frac{\|\hat{\theta}_{rdp} - \theta^*\|^2}{N} \leq \frac{C_3 Q_{rdp}(\theta^*)}{\epsilon^2} \tag{3.2}$$

where

$$Q_{rdp}(\theta^*) := \inf_{\theta \in \mathbb{R}^N} \left\{ \frac{\|\theta - \theta^*\|^2}{N} + \frac{k_{rdp}(\theta)(\max\{1, U^2\} \log^2(\max\{N, U\}) + \|\theta\|_\infty^2 \log N)}{N} \right\}.$$

Theorem 1 provides the generalization of the oracle risk bound in Theorem 2.1 in Chatterjee and Goswami (2019) to the quantile setting. We now list the differences of Theorem 1 with the oracle risk bound (Theorem 2.1 in Chatterjee and Goswami (2019)) known for the mean regression counterpart.

1. Theorem 2.1 in Chatterjee and Goswami (2019) requires that $Y - \theta^*$, the vector of errors, consists of i.i.d. mean zero Gaussian random variables. In contrast, Theorem 1 holds under Assumption 1 which does not require any tail decay assumptions for the distributions of the coordinates of the error vector (independence is still assumed). In particular, Assumption 1 allows error distributions with no moments as well like the Cauchy distribution.
2. The result in Chatterjee and Goswami (2019) is stronger in the sense that theirs is an upper bound in expectation, given as

$$\mathbb{E}\left(\frac{\|\hat{\theta} - \theta^*\|^2}{N}\right) \leq \inf_{\theta \in \mathbb{R}^N} \left\{ \frac{(1-\delta)}{(1+\delta)} \frac{\|\theta - \theta^*\|^2}{N} + \frac{C\sigma^2 k_{rdp}(\theta) \log N}{\delta(1-\delta)N} \right\}, \tag{3.3}$$

   for all $\delta \in (0, 1)$, and for some constant $C > 0$. Our result in Theorem 1 gives a tail probability inequality which does not ensure that $\frac{\|\hat{\theta}_{rdp} - \theta^*\|^2}{N}$ has a finite first moment. It does ensure however that

$$\frac{\|\hat{\theta}_{rdp} - \theta^*\|^2}{N} = \tilde{O}_{\mathbb{P}}(Q_{rdp}(\theta^*))$$

   where $\mathbb{P}$ refers to an appropriately defined sequence of probability distributions corresponding to denoising problems of increasing size.
3. In effect, the upper bound in Theorem 1 is only off by logarithmic factors compared to the upper bound in Theorem 2.1 in Chatterjee and Goswami (2019). Our bound in Theorem 1 contains some extra terms which are benign. The factor U should scale like $O(1)$ for any realistic error distribution sequence. The factor $\|\theta\|_\infty$ inside the infimum in the definition of $Q(\theta^*)$ essentially introduces another multiplicative factor of $\|\theta^*\|_\infty \leq U$.

Remark 1. *The choice of γ in Theorem 1 ensures that the QDCART estimator will be well behaved in the sense of the ℓ_∞ norm, see Lemma 1. Such a restriction on the size of the rectangles in the optimal partition is actually needed. Otherwise, the QDCART estimator can be arbitrarily large in some locations under the presence of heavy tailed errors. If one considers standard subGaussian type assumptions on the errors, then this restriction on the size of the rectangles can be removed, and U would need to be replaced with $U' > 0$ satisfying $U' = O(U + \sqrt{\log N})$, see Lemma 1.*

Remark 2. *Suppose that we construct our estimator for multiple quantile levels $\tau \in \Lambda \subset (0,1)$ where Λ is a finite set. Then, denoting by $\hat{\theta}_{rdp}(\tau)$ and $\theta^*(\tau)$ the QDCART and the true signal respectively for the quantile level τ, Theorem 1 implies that*

$$\frac{\|\hat{\theta}_{rdp}(\tau) - \theta^*(\tau)\|^2}{N} \leq \frac{C_3 Q_{rdp}(\theta^*(\tau))}{\epsilon^2} \quad \forall \tau \in \Lambda, \tag{3.4}$$

with probability at least $1 - C_2\epsilon|\Lambda|$. One might wonder whether it is possible to integrate over τ in Equation 3.4. Unfortunately, our main result in Theorem 1 is basically a $O_{\mathbb{P}}$ statement. Hence, in order to be able to integrate τ it would require us to develop new theoretical tools to those from the Appendix, as we believe solving this will require a new idea. Specifically, following our proof technique as in Appendix B, we can denote the population and empirical losses for τ as $M_\tau(\theta)$ and $\hat{M}_\tau(\theta)$ respectively. Then for a collection $\{t_\lambda\}$ of positive numers our goal becomes to derive an upper bound for the quantity

$$\mathbb{P}(\cup_{\tau\in\Lambda}\{\|\hat{\theta}_{rdp}(\tau) - \theta^*(\tau)\|^2 > t_\tau\}). \tag{3.5}$$

Hence, the peeling argument in Appendix B implies that we would need to give upper bounds for quantities of the form

$$A_j = \mathbb{P}\Bigg(\bigcup_{\tau\in\Lambda}\Bigg\{\sup_{\theta\in\Theta:\|\theta-\theta^*(\tau)\|^2\lesssim 2^j t_\tau^2} \big(M_\tau(\theta) + \hat{M}_\tau(\tilde{\theta}(\tau)) - \hat{M}_\tau(\theta) + \lambda k_{rdp}(\tilde{\theta}(\tau)) - \lambda k_{rdp}(\theta)\big) > 2^{j-1}t_\tau^2\Bigg\}\Bigg) \tag{3.6}$$

for $j \in \mathbb{N}$, and reference elements $\tilde{\theta}(\tau) \in \Theta$, and with Θ as in (C.3). For the case $|\Lambda| < \infty$, the conclusion in (3.4) follows by using union bound on (3.6) and then proceeding with the outline in Appendix B, bounding the terms for each τ separately. However, for $|\Lambda|$ that grows with n or that is infinite, the union bound approach would not allows to make (3.5) small. Thus, the main challenge becomes how to control (3.6) for general Λ, something that is not immediate from our current tools.

Remark 3. *A natural question that arises is whether the dependence on τ of the upper bound in Theorem 1 can be tracked. The answer to this is yes, in fact the only dependence through τ comes from the constant C_3 which satisfies $C_3 = O((L\underline{f})^{-1})$, with L and $\underline{f}$ as in Assumption 1. Hence, in principle, we can*

allow $\tau \to 0$ or $\tau \to 1$, but in that case the quantity $Q_{rdp}(\theta^*)$ would need to be inflated by $(L\underline{f})^{-1}$.

We now turn to the issue of computation. In this article we also give an algorithm to compute the QDCART estimator based on bottom up dynamic programming. This algorithm is similar to the original algorithm given in Donoho (1997) adapted to the quantile setting. We now state our computation result as a theorem.

Theorem 2. *There exists an absolute constant $C > 0$ (not depending on d, n) such that the computational complexity, i.e. the number of elementary operations involved in the computation of the QDCART estimator in d dimensions is bounded by $C\big(N(\log n)^d + d\, 2^d N\big)$.*

The description of the algorithm and the proof of its computational complexity are given in Section 5.

3.1. Implications for bounded variation signals

It was shown in Donoho (1997) and Chatterjee and Goswami (2019) that an oracle risk bound of the type shown in Theorem 1 implies minimax rate optimality (up to log factors) for other function classes of interest as well. We now proceed to discuss consequences of Theorem 1 for the class $\mathcal{BV}_{d,n}(V)$ of bounded variation signals. This class of signals is defined as

$$\mathcal{BV}_{d,n}(V) := \left\{\theta \in \mathbb{R}^{L_{d,n}} : \mathrm{TV}(\theta) \le V\right\},$$

where

$$\mathrm{TV}(\theta) := \sum_{(i,j)\in E_{d,n}} |\theta_i - \theta_j|,$$

and $E_{d,n}$ is the edge set of the graph $L_{d,n}$.

The class of signals $\mathcal{BV}_{d,n}(V)$ is rich enough to contain signals that are smooth in certain regions of their domain but discontinuous in other regions. The problem of estimation of a signal in the class $\mathcal{BV}_{d,n}(V)$ has attracted a lot of attention in the statistics literature, see for instance Mammen and van de Geer (1997); Tibshirani (2014); Sadhanala et al. (2016); Hutter and Rigollet (2016); Padilla et al. (2018); Chatterjee and Goswami (2021); Ortelli and van de Geer (2019); Guntuboyina et al. (2020).

We arrive at the next corollary by combining Theorem 1 with existing approximation theoretic results shown in Chatterjee and Goswami (2019) (see Proposition 8.9 and Theorem 4.2 there)

Corollary 1. *For any $\theta^* \in \mathcal{BV}_{d,n}(V)$, there exists a constant $C > 0$ only*

depending on the dimension d such that

$$Q_{rdp}(\theta^*) \leq \begin{cases} C\left(\frac{V^{2/3}\max\{1,U^2\}\log^{5/3}(\max\{N,U\})}{N^{2/3}} + \frac{\max\{1,U^2\}\log^2\max\{N,U\}}{N}\right) \\ \quad \text{if } d = 1 \\ C\left(\frac{V\max\{1,U^2\}\log^2(\max\{N,U\})}{N} + \frac{\max\{1,U^2\}\log^2\max\{N,U\}}{N}\right) \\ \quad \text{if } d > 1. \end{cases}$$

Therefore, under the same assumptions and the choice of λ and γ in Theorem 1, the same probability tail bound as in (3.2) holds for any $\theta^* \in \mathcal{BV}_{d,n}(V)$ with $Q_{rdp}(\theta^*)$ replaced by the bound above.

The rates implied by Corollary 1 are minimax optimal, save for logarithmic factors, in the class $\mathcal{BV}_{d,n}(V)$, see the discussion in Tibshirani (2014) for the case $d = 1$ and the corresponding one in Hutter and Rigollet (2016), Sadhanala et al. (2016) for the case $d > 1$. It was shown in Theorem 5.1 from Chatterjee and Goswami (2019) that the mean regression version of Dyadic CART is minimax rate optimal (up to log factors) in the class $\mathcal{BV}_{d,n}(V)$. Corollary 1 can be seen as an extension of this result to the quantile setting which holds under much weaker tail decay conditions.

3.2. Implications for piecewise constant signals

We now discuss consequences of Theorem 1 for the class of piecewise constant signals in dimensions $d = 1$ and $d = 2$. Towards that end, given $\theta \in \mathbb{R}^{L_{d,n}}$, we define $k(\theta)$ *as the size of the smallest rectangular partition* Π *of* $L_{d,n}$ *such that* θ *is constant in each rectangle of* Π. By construction, $k(\theta) \leq k_{rdp}(\theta)$ for all $\theta \in \mathbb{R}^{L_{d,n}}$. Furthermore, Proposition 3.9 in Chatterjee and Goswami (2019) shows that there exists an absolute constant $C > 0$ such that for all $\theta \in \mathbb{R}^{L_{d,n}}$ it holds that

$$k_{rdp}(\theta) \leq Ck(\theta)\log\left(\frac{en}{k(\theta)}\right) \tag{3.7}$$

if $d = 1$ and

$$k_{rdp}(\theta) \leq C(\log n)^2 k(\theta) \tag{3.8}$$

if $d = 2$.

Combining Theorem 1 with (3.7) and (3.8) we immediately obtain our next corollary

Corollary 2. *For any $\theta^* \in \mathbb{R}^{L_{d,n}}$, there exists a constant $C > 0$ only depending on the dimension d such that*

$$Q_{rdp}(\theta^*) \leq \begin{cases} C\left(\frac{k(\theta^*)\max\{1,U^2\}\log^2(\max\{N,U\})\log(N/k(\theta^*))}{N}\right) & \text{if } d = 1 \\ C\left(\frac{k(\theta^*)\max\{1,U^2\}\log^2(\max\{N,U\})\log^2 N}{N}\right) & \text{if } d = 2. \end{cases}$$

Therefore, under the same assumptions and the choice of λ and γ in Theorem 1, the same probability tail bound as in (3.2) holds for any $\theta^* \in \mathbb{R}^{L_{d,n}}$ with $Q_{rdp}(\theta^*)$ replaced by the bound above.

Notice that in Corollary 2, the resulting rate implied is $\tilde{O}(k(\theta^*)/N)$ which is the usual parametric rate of estimation for a signal θ^* consisting of $k(\theta^*)$ pieces if one knows the locations of the end points of the constant pieces of θ^*. Here, of course the QDCART estimator does not know the true partition corresponding to the true signal. Corollary 2 can be seen as an extension of Corollary 3.10 in Chatterjee and Goswami (2019) to the quantile setting which holds even under heavy tailed error distributions.

Remark 4. *The situation when $d > 2$ is more difficult as versions of (3.7) and (3.8) are not known to hold in higher dimensions than 2. We refer the reader to Chatterjee and Goswami (2019) where this issue has been thoroughly discussed. We prefer therefore to just state our results for dimensions $d \leq 2$.*

Remark 5. *All our theoretical guarantees are in the regime when d is held fixed and n is growing. Our estimator and our results are practically useful when $d \in \{1, 2, 3\}$.*

4. Discussion

4.1. Comparison with quantile total variation denoising

We have shown that QDCART is computable in near linear time and enjoys attractive statistical properties. We do not compare QDCART with corresponding mean regression estimators simply because under heavy tails, the mean regression estimators perform poorly while our results continue to hold; see the simulations section in Padilla and Chatterjee (2020). Therefore, comparison is appropriate with other quantile regression estimators. We believe that the most natural competitor to QDCART is the Quantile Total Variation Denoising (QTVD) estimator, proposed and studied in Padilla and Chatterjee (2020). Actually, there are two versions of QTVD, the so called constrained and the penalized version. In the univariate case, Padilla and Chatterjee (2020) refers to the constrained version of this estimator as constrained quantile fused lasso (CQFL), and to the penalized version as penalized quantile fused lasso (PQFL). Both of these estimators were analyzed in Padilla and Chatterjee (2020).

In comparing Corollaries 1 and 2 with the existing results known for CQFL and PQFL estimators, we make the following points.

1. The results proven in Padilla and Chatterjee (2020) for CQFL and PQFL only need the lower bound portion in Assumption 1 while Corollaries 1 and 2 require also the upper bound in Assumption 1 to hold. However, this is a very mild condition that leads to a guarantee for QDCART in terms of the mean squared error, a stronger (and a much cleaner) result than those for CQFL and PQFL which are based on the loss $\Delta_N^2(\cdot)$ defined as

$$\Delta_N^2(v) := \frac{1}{N} \sum_{i \in L_{d,n}} \min\{|v_i|, v_i^2\}, \ \forall v \in \mathbb{R}^{L_{d,n}}. \tag{4.1}$$

Results under the squared error loss are not yet known for the CQFL and PQFL estimators. The main reason for this is that the localization result described in Step 1 in Section B is not available for CQFL and PQFL.

2. The dependence on the total variation of the true signal V in both Corollaries 1 and 2 are optimal in the sense that they match the right dependence known in the mean regression case; see Sadhanala et al. (2016) for lower bounds on bounded variation signal classes. The results for the CQFL and PQFL estimators in Padilla and Chatterjee (2020) seem to have sub optimal dependence on V. Thus, to the best of our knowledge, our results on QDCART are the first quantile regression based estimators which enjoy minimax rate optimality with respect to both the sample size N and the total variation of the unknown signal V.
3. Our results for QDCART depend on the quantity U something that is not the case for QTVD. However, in any realistic setting, we would have $U = O(1)$.
4. Corollary 2 gives a near parametric rate of convergence for piecewise constant signals. In the univariate case, the corresponding result is known for the CQFL and PQFL in Theorems 2 and 4 from Padilla and Chatterjee (2020). However, these results need the true signal to satsify a *minimal spacing condition* which is not needed for QDCART. This is potentially a significant advantage of QDCART over QTVD, even in $d = 1$, as far as attaining adaptively optimal rates of convergence for piecewise constant signals is concerned.
5. In the mean regression problem, it is known that when $d = 2$, the TVD estimator cannot attain near parametric rates of convergence for a rectangular piecewise constant signal; see Theorem 2.3 in Chatterjee and Goswami (2021). Therefore, it is expected that the QTVD estimator would also not be the best tool for estimating rectangular piecewise constant signals. On the other hand, the QDCART estimator does attain the $\tilde{O}(\frac{k(\theta^*)}{N})$ rate in 2 dimensions and seems to be the right tool for estimating rectangular piecewise constant signals as well.
6. The QDCART estimator is computable in $\tilde{O}(N)$ time in any dimensions. In contrast, it is unknown and unlikely that the QTVD estimator is computable in $\tilde{O}(N)$ time in dimension larger than 1.

Remark 6. *The last three points above show that the QDCART estimator is a computationally faster alternative to the QTVD estimator while also enjoying some statistical advantages. We perform numerical experiments to further compare finite sample performance of QDCART and QTVD estimators in Section 6.*

4.2. Background and related literature

Our work in this paper falls under the scope of nonparametric quantile regression. We now briefly review some classical work on nonparametric quantile regression. In the context of median regression some early works include Utreras (1981), Cox (1983), and Eubank (1988). Koenker et al. (1994) proposed

one dimensional quantile smoothing splines. These estimators were studied in He and Shi (1994) under the assumption that the quantile function is Hölder continuous.

Other related quantile nonparametric estimators include the bivariate quantile smoothing splines studied in He et al. (1998), the tree based estimator from Chaudhuri and Loh (2002), the quantile random forest proposed by Meinshausen and Ridgeway (2006), and the generalized random forest from Athey et al. (2019). van de Geer (2003) developed general bounds for nonparametric quantile regression. Brown et al. (2008) constructed a wavelet-based estimator for median regression with Besov functions. Recently, Ye and Padilla (2021) developed the k-nearest neighbour quantile fused lasso approach and Padilla et al. (2020) studied quantile regression with ReLU networks.

4.3. Future directions

There are different research directions that we leave for future work. A natural extension is to consider piecewise polynomial structures in the estimator, similarly to Chatterjee and Goswami (2019). However, we are currently unaware of how to extend our theory to such a setting. The main bottleneck is that given a fixed sub rectangle of $L_{d,n}$ we do not know how to obtain an ℓ_∞ upper bound when fitting quantile regression constrained to the class of polynomials of degree $r > 0$. When $r = 0$ the latter can be done as in Corollary 1. This a crucial ingredient in our proof that we do not know how to handle when dealing with higher order piecewise polynomial signals.

Moreover, it would be worthwhile to mention here that all our theoretical results hold under a theoretical choice of the tuning parameters. In our experiments, following Yu and Moyeed (2001), we choose the tuning parameters by Bayesian information criterion (BIC) for quantile regression. It would be interesting to provide theoretical guarantees for an estimator which chooses the tuning parameters in a data driven way, for example, by some form of cross validation.

5. Computation of the QDCART estimator

The goal of this section is to develop a computationally efficient algorithm for the QDCART estimator defined in(2.2). In doing so, our construction will imply the conclusion of Theorem 2. This algorithm is an adaptation of the original algorithm given in Donoho (1997) (also see Lemma 1 in Chatterjee and Goswami (2019)) to the quantile setting.

We have to solve the discrete optimization problem in (2.3). Let us first see how we can solve the discrete optimization problem in (2.1) where there are no constraints on the size of the rectangles.

In order to study the optimization problem (2.1), we first define a corresponding subproblem for any rectangle R. For any rectangle $R \subset L_{d,n}$ and a partition $\Pi \in \mathcal{P}_{rdp}(L_{d,n})$, we let $\Pi(R) := \{A \cap R : A \cap R \neq \emptyset,\ A \in \Pi\}$ be the

partition induced by Π in R. We then let $\mathcal{P}_{rdp}(R) := \{\Pi(R) : \Pi \in \mathcal{P}_{rdp}(L_{d,n})\}$. In words, $\mathcal{P}_{rdp}(R)$ is the set of recursive dyadic partitions of the rectangle R.

Then we define the following subproblem and define its optimal value as

$$\mathrm{OPT}(R) := \min_{\Pi \in \mathcal{P}_{rdp}(R)} \left\{ \sum_{i \in R} \rho_\tau(y_i - (O_{\Pi,\tau}(y))_i) + \lambda |\Pi| \right\}.$$

Clearly, $\mathrm{OPT}(L_{d,n})$ is the optimal value of the objective function associated with QDCART. The basic idea is to be able to solve *smaller* subproblems as above and build these smaller solutions to solve the full optimization problem. The following *dynamic programming* relation allows us to build up from bottom up

$$\mathrm{OPT}(R) := \min_{R_1,R_2 \text{ dyadic split of} R} \left\{ \mathrm{OPT}(R_1) + \mathrm{OPT}(R_2), \sum_{i \in R} \rho_\tau(y_i - q_\tau(y_R)) + \lambda \right\},$$

where by saying that "R_1 and R_2 dyadic split of R" we mean that R_1 and R_2 were obtained after performing a dyadic split of R. The above relation follows because of the separable nature of our optimization objective and the second term inside the minimum above corresponds to not splitting R at all.

We now proceed visiting dyadic rectangles bottom-up according to the length of R. The length of R is defined as the sum of the lengths of the sides of the rectangles. We will start from the minimum possible length d all the way to nd. Our goal is to store $\mathrm{OPT}(R)$ and $\mathrm{SPLIT}(R)$ for each dyadic rectangle R, where $\mathrm{SPLIT}(R)$ indicates the optimal split for rectangle R. Note that the total number of possible splits is d (one for each dimension) and thus $\mathrm{SPLIT}(R)$ can be represented by a single integer within the set $[d]$.

For each dyadic rectangle R, let us denote

$$\mathrm{SQL}(R) := \sum_{i \in R} \rho_\tau(y_i - q_\tau(y_R)). \tag{5.1}$$

where SQL stands for sum of quantile loss and $q_\tau(y_R)$ is an empirical τ quantile of the set of observations in y_R. Assume that we have succesfully computed $\mathrm{SQL}(R)$ for each dyadic rectangle. Then, at each dyadic rectangle R, to compute $\mathrm{OPT}(R)$ we have to compute the sum $\mathrm{OPT}(R_1) + \mathrm{OPT}(R_2)$ for each possible non trivial dyadic split of R into R_1, R_2 and then compute the minimum of $d+1$ numbers. Once $\mathrm{OPT}(R)$ is computed, $\mathrm{SPLIT}(R)$ is also automatically calculated when we are computing the minimum of these $d+1$ numbers.

Note that since we are visiting dyadic rectangles bottom up, we have already computed $\mathrm{OPT}(R')$ for all sub rectangles $R' \subset_{\neq} R$. Therefore, the computation required for computing $\mathrm{OPT}(R)$ is $d+1$. The total number of dyadic rectangles is at most $2^d N$. Therefore, the total computation required to compute $\mathrm{OPT}(R)$ for all dyadic rectangles R is at most $(d+1)2^d N$.

Now we proceed to explain how to compute $\mathrm{SQL}(R)$ for each dyadic rectangle R. We again do this by a bottom up scheme by visiting dyadic rectangles

according to their length (small to large). Our aim here is to compute a sorted list of observations within each dyadic rectangle R. We do this bottom up. For each dyadic rectangle R we consider a particular dyadic split R_1 and R_2 which is obtained by dyadically splitting (in dictionary order) the first coordinate of R. In the case the first coordinate is a singleton, we use the second coordinate to split and so on. Now on our bottom up visits, we iteratively compute sorted lists for these dyadic rectangles. For instance, for a given dyadic rectangle R, we take its corresponding dyadic split (in dictionary order) into R_1, R_2. Now we have already created a sorted list for R_1 and R_2. To create a sorted list for R we just need to merge two sorted lists. We can do this by the standard merge sort algorithm. The computation required at this step is $O(|R_1|) + O(|R_2|) = O(|R|)$. Once we are able to construct the sorted list of observations within R, now $\mathrm{SQL}(R)$ can be readily computed in $O(|R|)$ time.

Now consider a dyadic rectangle R of a given size $2^{i_1} \times \ldots \times 2^{i_d}$. The total number of dyadic rectangles of this size $2^{i_1} \times \ldots \times 2^{i_d}$ is at most

$$\frac{n}{2^{i_1}} \times \ldots \frac{n}{2^{i_d}}.$$

Therefore, the total computational work needed to compute $\mathrm{SQL}(R)$ for all dyadic rectangles R with this given size is simply

$$O(2^{i_1} \times \ldots \times 2^{i_d} \times \frac{n}{2^{i_1}} \times \ldots \frac{n}{2^{i_d}}) = O(n^d).$$

Now note that the total number of distinct sizes $2^{i_1} \times \ldots \times 2^{i_d}$ is at most $(\log n)^d$. Therefore, the total computational work needed to compute $\mathrm{SQL}(R)$ for all dyadic rectangles of all sizes with our bottom up scheme is $O(N(\log n)^d)$.

Finally, we see that the total computation required to compute $\mathrm{OPT}(R)$ and $\mathrm{SPLIT}(R)$ for all dyadic rectangles R is $O(N(\log n)^d + d2^d N)$. After $\mathrm{OPT}(R)$ and $\mathrm{SPLIT}(R)$ have been constructed, we can find the optimal partition by going top-down. This would be a lower order computation.

Based on the discussion above, it is not hard to see that to compute (2.3) we just have to modify the above algorithm slightly. We do not need to compute $\mathrm{OPT}(R)$ and $\mathrm{SPLIT}(R)$ for rectangles R with $|R| < \gamma$. Thus, when we visit dyadic rectangles bottom up according to its size, we just visit the *feasible* rectangles. Also, for a given rectangle R, to compute $\mathrm{OPT}(R)$ we now have to compute the sum $\mathrm{OPT}(R_1) + \mathrm{OPT}(R_2)$ for each possible non trivial dyadic split of R into R_1, R_2 where R_1, R_2 are both *feasible* and then compute the minimum of these numbers and $\sum_{i \in R} \rho_\tau(y_i - q_\tau(y_R))$.

Remark 7. *Notice that if the goal is to find (2.1) for multiple $\tau \in \Lambda \subset (0,1)$, the bottom up calculations described before can be done simultaneously for each τ. Specifically, we can compute* $\mathrm{OPT}(R)$ *and* $\mathrm{SPLIT}(R)$ *for all rectangles R and all values of τ exploiting the sorted arrays for each rectangle R.*

6. Experiments

6.1. Comparisons in 1d

We now proceed to evaluate the performance of QDCART in the 1d setting. In our simulations we consider as benchmarks the penalized quantile fused lasso (PQFL) proposed in Brantley et al. (2019) and studied in Padilla and Chatterjee (2020), and the univariate mean regression DCART method from Donoho (1997). For our evaluations in this subsection, for QDCART and DCART we consider a grid of 25 values of λ given as $\{2^{-2}, 2^{-1.75}, \ldots, 2^{4}\}$ and we set $\gamma = 8$, a choice that we find to work well in practice and the results of this section are not sensitive to the choice of γ. As for PQFL, we take λ such that $\log \lambda \in \{1 + \frac{j(7.5-1)}{99} : j \in \{0, 1, \ldots, 99\}\}$. Then, for each method and choice of tuning parameter we calculate the average mean squared error, averaging over 100 data sets generated from different scenarios and with $n \in \{512, 1024\}$. For each method and each scenario we then report the optimal MSE. The only remaining ingredient is to explain the different scenarios for generating data that we consider. These are described next.

For each scenario, we generate the data $y \in \mathbb{R}^{L_{1,n}}$ as $y_i = \theta_i^* + \epsilon_i$, for $i \in L_{1,n}$ and some $\theta^*, \epsilon \in \mathbb{R}^{L_{1,n}}$. We now explain the constructions of θ^* and ϵ for the different scenarios.

Scenario 1. (Large Segments). In this case $\theta^* \in \mathbb{R}^{L_{1,n}}$ satisfies

$$\theta_i^* = \begin{cases} 1 & \text{if } i \in [\lfloor n/5 \rfloor + 1, 2\lfloor n/5 \rfloor] \cup [3\lfloor n/5 \rfloor + 1, n] \\ 0 & \text{otherwise,} \end{cases}$$

and we generate $\epsilon_i \overset{\text{ind}}{\sim} t(2.5)$ where $t(2.5)$ denotes the t-distribution with 2.5 degrees of freedom.

Scenario 2. (Large and Small Segments). We generate ϵ as in Scenario 1 and set

$$\theta_i^* = \begin{cases} 1 & \text{if } i \in [\lfloor n/3 \rfloor + 1, \lfloor n/3 \rfloor + \lfloor n/32 \rfloor] \cup [\lfloor n/3 \rfloor + 2\lfloor n/32 \rfloor + 1, \\ & \quad \lfloor n/3 \rfloor + 3\lfloor n/32 \rfloor] \cup [\lfloor n/3 \rfloor + 4\lfloor n/32 \rfloor + 1, n], \\ 0 & \text{otherwise.} \end{cases}$$

Scenario 3. (Large Segments and Cauchy Errors). We take $\theta^* \in \mathbb{R}^{L_{1,n}}$ as in Scenario 1 and generate $\epsilon_i \overset{\text{ind}}{\sim} \text{Cauchy}(0, 1)$.

Scenario 4. (Large and Small Segments, and Heteroscedastic Errors). The vector θ^* is the same as in Scenario 2 and ϵ satisfies for all i that

$$\epsilon_i = \nu_i \times \sqrt{\frac{2i}{n} + 1},$$

where $\nu_i \overset{\text{ind}}{\sim} \text{N}(0, 1)$.

A visualization of data generated under each scenario is given in Figure 3. The results of our comparisons are given in Table 1. Overall, we can see that

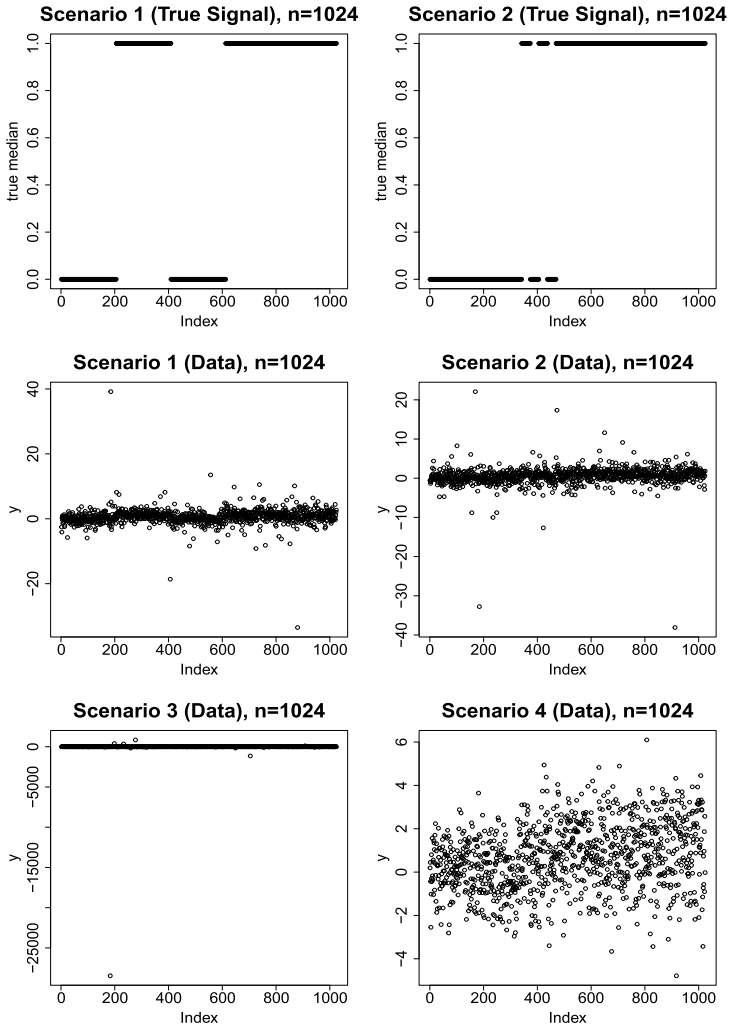

FIG 3. *True signal θ^* and instances of data generated under each of the scenarios.*

TABLE 1
Average mean squared error $\frac{1}{n}\sum_{i=1}^{n}(\theta_i^* - \hat{\theta}_i)^2$, averaging over 100 Monte carlo simulations for the different methods considered. Captions are described in the text.

n	Scenario	PQFL	QDCART	DCART	n	Scenario	PQFL	QDCART	DCART
512	1	0.124	**0.094**	3.08	512	3	**0.177**	0.252	249054.2
1024	1	**0.047**	0.066	2.52	1024	3	**0.118**	0.249	104763.3
512	2	0.084	**0.063**	3.17	512	4	0.090	**0.070**	0.114
1024	2	0.064	**0.047**	2.99	1024	4	0.077	**0.054**	0.106

the QDCART estimator is competitive against PQFL. In Scenario 2, where some of the constant pieces of the true signal are very small, we see that the QDCART estimator performs better. This is in agreement with Corollary 2 where no minimum length assumption is needed for the QDCART to attain near parametric rates. Observe that the mean regression DCART estimator performs poorly under heavy tailed scenarios.

6.2. Comparisons in 2d

We now proceed to evaluate the performance of QDCART for 2d grid graphs and use DCART and QTVD (PQFL) as benchmarks. For our experiments in this subsection the tuning parameter λ for QDCART and DCART is taken such that $\log_{10}(\lambda)$ is in the set $\{-1 + \frac{6.5j}{59} : j \in \{0, 1, \ldots, 59\}\}$. As for the tuning parameter λ for QTVD we take it such that $\log_2(\lambda)$ is in the set $\{-1 + \frac{7j}{19} : j \in \{0, 1, \ldots, 19\}\}$. As before, for each method and choice of tuning parameter we calculate the average mean squared error, averaging over 100 data sets generated from different scenarios. We set $d = 2$ and $n \in \{64, 128, 256\}$. For each method and each scenario we then report the optimal MSE. Next we describe the different generative models, where in each case the data are generated as $y_{i,j} = \theta^*_{i,j} + \epsilon_{i,j}$ where $\epsilon_{i,j} \overset{\text{ind}}{\sim} t(2.5)$ for $i, j \in \{1, \ldots, n\}$ and with $\theta^* \in \mathbb{R}^{n \times n}$.

Scenario 5. We set

$$\theta^*_{i,j} = \begin{cases} 1 & \text{if } n/5 < i < 3n/5 \text{ and } n/5 < j < 3n/5, \\ 0 & \text{otherwise.} \end{cases}$$

Scenario 6. Now we take θ^* satisfying

$$\theta^*_{i,j} = \begin{cases} 1 & \text{if } (i - n/4)^2 + (j - n/4)^2 < (n/5)^2, \\ -1 & \text{if } (i - 3n/4)^2 + (j - 3n/4)^2 < (n/5)^2, \\ 0 & \text{otherwise.} \end{cases}$$

Scenario 7. For this model we let

$$\theta^*_{i,j} = \begin{cases} 1 & \text{if } n/4 < i < 3n/4 \text{ and } n/4 < j < n/4 + n/8, \\ 1 & \text{if } n/2 + n/8 < i < 3n/4 \text{ and } n/4 + n/8 \le j < 3n/4, \\ -1 & \text{if } i > 6n/8 \text{ and } j > 6n/8, \\ 0 & \text{otherwise.} \end{cases}$$

Scenario 8. This is the same as Scenario 7, but unlike the previous scenarios, we now generate $\epsilon_{i,j} \overset{\text{ind}}{\sim} N(0, 1)$.

A visualization of the data generated under Scenarios 5–7 is provided in Figure 4. There we can see that QDCART can be competitive against QTVD.

A more comprehensive evaluation of performance comparisons is provided in Table 2. In Scenario 6 where the level sets are non-rectangular, QTVD seems to do better than QDCART. In Scenario 7, however, QDCART performs slightly

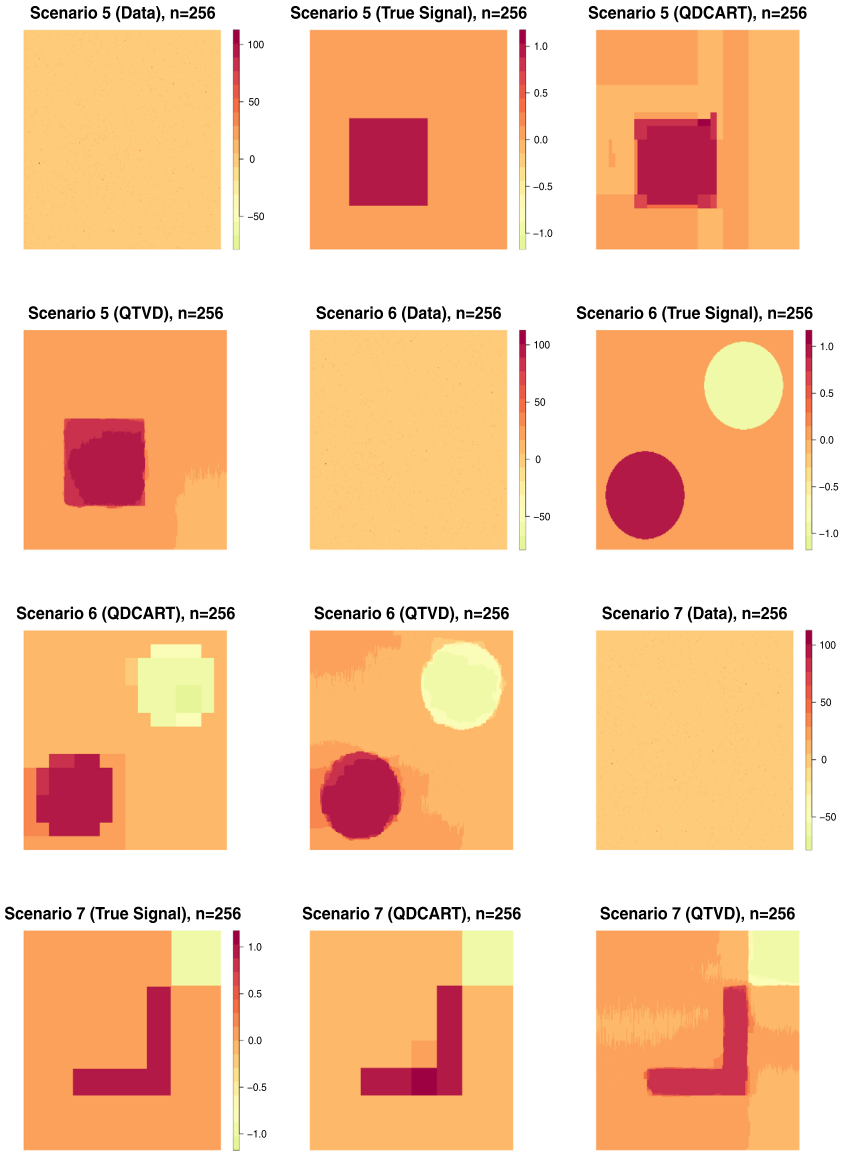

FIG 4. *From left to right, an instance of data, true signal and estimates for Scenarios 5 to 7.*

better. We believe this is because the level set can be well represented by a dyadic partition. We reiterate here that a potential practical advantage of QDCART over QTVD in an image denoising setting is the fact that the QDCART estimator can be computed in near linear time.QTVD algorithm, being the solution of a convex optimization program, is slower and no near linear time algorithm for solving it is known.

TABLE 2
Average mean squared error $\frac{1}{n}\sum_{i=1}^{n}(\theta_i^* - \hat{\theta}_i)^2$*, averaging over 100 Monte carlo simulations for the different methods considered and with data generated from Scenarios 5–7. Captions are described in the text.*

n	Scenario	QTVD	QDCART	DCART	Scenario	QTVD	QDCART	DCART
64	5	**0.030**	0.048	0.139	6	**0.057**	0.096	0.250
128	5	**0.013**	0.021	0.134	6	**0.023**	0.035	0.252
256	5	**0.004**	0.009	0.133	6	**0.011**	0.026	0.251
n	Scenario	QTVD	QDCART	DCART	Scenario	QTVD	QDCART	DCART
64	7	0.060	**0.033**	0.157	8	0.048	0.021	**0.016**
128	7	0.022	**0.009**	0.163	8	0.018	0.008	**0.005**
256	7	0.009	**0.004**	0.166	8	0.007	0.004	**0.001**

Finally, we also see in Table 2, that under Scenario 8, DCART is the best method. This is not surprising since the error terms are Gaussian distributed.

6.3. Ion channels data

We conclude our experiments section with a real data example involving ion channels data. Ion channels are a class of proteins expressed by all cells that create pathways for ions to pass through the cell membrane. As explained in Jula Vanegas et al. (2021), over time, the ion channel changes its gating behavior by closing and reopening its pore which leads to a piecewise constant current flow structure. The original data that we use was produced by the Steinem Lab (Institute of Organic and Biomolecular Chemistry, University of Gottingen), and it was recently analyzed by Jula Vanegas et al. (2021). It consists of a single ion channel of the bacterial porin PorB, a bacterium that plays a role in the pathogenicity of Neisseria gonorrhoeae. The original data is 600000 time instances. For our comparisons we focus on a portion of length 32511 and construct a subsampled vector $y \in \mathbb{R}^{2048}$. Thus, our resulting signal is similar to that in Cappello et al. (2021). A depiction of the data is shown in Figure 5.

Given the signal $y \in \mathbb{R}^{2048}$, we fit both PQFL and QDCART with values $\tau \in \{0.1, 0.5, 0.9\}$. For PQFL we consider values of the penalty parameter λ such that $\log\lambda$ is in the set $\{1 + \frac{j6.5}{99} : j \in \{0, \ldots, 99\}\}$. As for QDCART we take λ such that $\log_2\lambda$ is in $\{-2 + \frac{7j}{25} : j = 0, \ldots, 25\}$. Then for both PQFL and QDCART we choose the tuning parameter that minimizes the BIC for quantile regression criteria from Yu and Moyeed (2001) given as

$$\text{BIC} := \frac{2}{\sigma}\sum_{i=1}^{n}\rho_\tau(y_i - \hat{\theta}_i) + v\log n,$$

where as in Brantley et al. (2019) and Ye and Padilla (2021) we take $\sigma = \frac{1-|1-2\tau|}{2}$, and where v is the estimated degrees of freedom. Motivated by Tibshirani and Taylor (2012), we let

$$v = \left|\left\{j : |\hat{\theta}_j - \hat{\theta}_{j+1}| > 10^{-3}\right\}\right|.$$

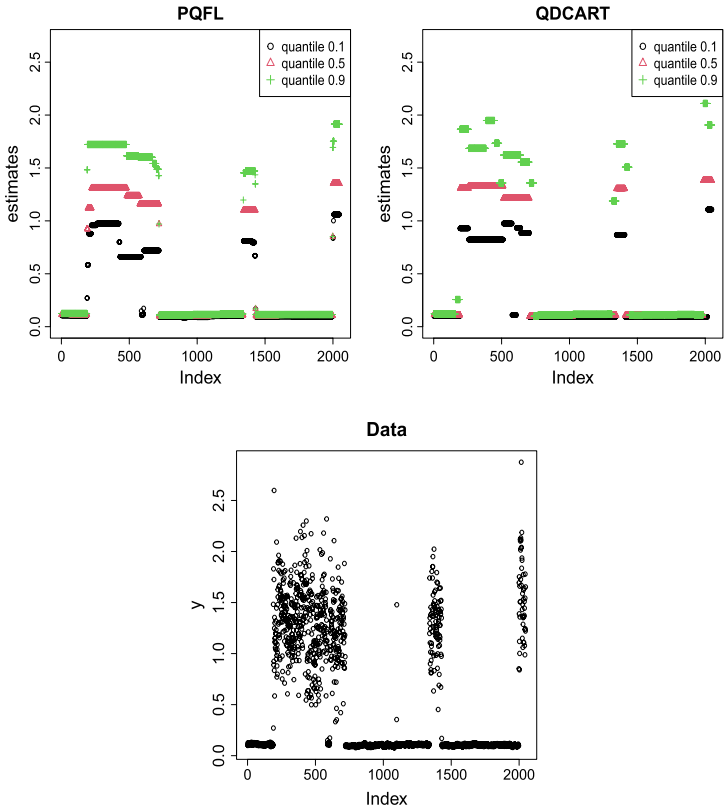

FIG 5. *The first two panels show the estimated PQFL and QDCART for* $\tau \in \{0.1, 0.5, 0.9\}$ *when using the ion data. The last panel then shows the ion raw data.*

With the above choice of tuning parameter for both PQFL and QDCART, we compute the estimates which are displayed in Figure 5. There, we can see that the estimators are roughly similar, validating our theoretical findings that QDCART and PQFL have similar statistical properties.

Finally, in order to provide a clearer quantitative comparison, we proceed as follows. We randomly choose 50% of the entries of the signal described above and we use this as training. Then we use the remaining 50% of the data as testing, and for each coordinate in the test set we make a prediction based on the closest coordinate in the training set. With this in hand, for each competing method, we compute $\text{prop}_{0.5}$, the proportion of the test samples that are below its predicted median. We also compute $\text{cov}_{80\%}$, the proportion of samples in the test set that are between their predicted 0.1 and 0.9 quantiles. The quantities $\text{prop}_{0.5}$ and $\text{cov}_{80\%}$ are then averaged over 100 repetitions and reported for PQFL and QDCART. For QDCART we obtain the values $\text{prop}_{0.5} = 0.502$ and $\text{cov}_{80\%} = 0.781$, whereas for PQFL we obtain $\text{prop}_{0.5} = 0.502$ and $\text{cov}_{80\%} = 0.772$. These results suggest that QDCART provides ever so slightly better prediction intervals than PQFL.

Appendix A: Quantile ORT estimator

The ORT estimator, introduced in Chatterjee and Goswami (2019) is a variant of the Dyadic CART estimator which enjoys better statistical risk guarantees in general but has significantly slower computational complexity; see Lemma 1 in Chatterjee and Goswami (2019). Just as we have proposed QDCART, it is natural to extend the optimal regression tree (ORT) estimator to the quantile setting as well. This leads us to define the quantile optimal regression tree (QORT) estimator. Before giving the definition of QORT, we need to introduce some additional notation.

Given a rectangle $R = \prod_{j=1}^{d}[a_j, b_j] \subset L_{d,n}$, a hierarchical split consists of choosing a coordinate $j \in \{1, \ldots, d\}$ and then constructing the rectangles R_1 and R_2 with $R = R_1 \cup R_2$, $R_1 \cap R_2 = \emptyset$ and

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

with $a_j \leq l \leq b_j$ and $l \in \mathbb{Z}_+$. Thus, the difference of a hierarchical split with a dyadic split is that the former is not restricted to split an interval only at the midpoint. Starting from $L_{d,n}$, one can keep on performing hierarchical splits recursively, creating refined partitions. A *hierarchical partition/decision tree* is any partition reachable by such successive hierarchical splits. Note that this is the usual definition of a decision tree except we are carrying out everything on the lattice $L_{d,n}$. We denote by $\mathcal{P}_{tree}(L_{d,n})$ the set of hierarchical partitions of $L_{d,n}$.

Given $\theta \in \mathbb{R}^{L_{d,n}}$, we denote by $k_{tree}(\theta)$ *the smallest number of elements of any hierarchical partition in which* θ *is piecewise constant*. It is clear that for any $\theta \in \mathbb{R}^{L_{d,n}}$ we must have

$$k(\theta) \leq k_{tree}(\theta) \leq k_{rdp}(\theta).$$

Armed with the notation above, we can now define the estimator

$$\hat{\theta}_{tree} = O_{\hat{\Pi}_{tree},\tau}(y) \tag{A.1}$$

where

$$\hat{\Pi}_{tree} := \underset{\Pi \in \mathcal{P}_{tree}(L_{d,n}) : |R| \geq \gamma \ \forall R \in \Pi}{\arg\min} \left\{ \sum_{i \in L_{d,n}} \rho_\tau(y_i - (O_{\Pi,\tau}(y))_i) + \lambda |\Pi| \right\} \tag{A.2}$$

for tuning parameters $\lambda, \gamma > 0$. By construction, $\hat{\theta}_{tree}$ is the quantile version of the ORT estimator proposed and studied in Chatterjee and Goswami (2019).

With the notation above in hand, we are now ready to present our main result for the QORT estimator.

Theorem 3. *Define for any $\theta \in \mathbb{R}^{L_{d,n}}$, the quantity*

$$Q_{tree}(\theta^*) := \inf_{\theta \in \mathbb{R}^N} \left\{ \frac{\|\theta - \theta^*\|^2}{N} + \frac{k_{tree}(\theta)(\max\{1, U^2\} \log^2(\max\{N, U\}) + \|\theta\|_\infty^2 \log N)}{N} \right\}$$

Under the same assumptions and the choice of λ and γ in Theorem 1, the same probability tail bound as in (3.2) holds for any $\theta^* \in \mathbb{R}^{L_{d,n}}$ with one difference; $Q_{rdp}(\theta^*)$ is replaced by $Q_{tree}(\theta^*)$.

Remark 8. *The above theorem basically says that $\frac{1}{N}\|\hat{\theta}_{tree} - \theta^*\|^2 = O_{\mathbb{P}}(Q_{tree}(\theta^*))$. This is in general a better bound than saying $\frac{1}{N}\|\hat{\theta}_{tree} - \theta^*\|^2 = O_{\mathbb{P}}(Q_{rdp}(\theta^*))$ because $k_{tree}(\theta^*) \leq k_{rdp}(\theta^*)$.*

Theorem 3 generalizes to the quantile setting the general risk bound proven in Chatterjee and Goswami (2019) for the ORT estimator. It is clear that the implications presented for bounded variation and piecewise constant function classes continue to hold for the QORT estimator as well. However, the QORT estimator would have significantly worse computational complexity than the QDCART estimator which is why we focus more on the QDCART estimator in this paper. It should be possible to provide an algorithm demonstrating a computational complexity result scaling like $\tilde{O}(N^{2+1/d})$ for the QORT estimator, analogous to Theorem 2. We do not carry this due to space considerations.

Appendix B: Proof technique

Our proof follows a M-estimation approach by viewing the QDCART estimator as a penalized M estimator. This M estimation viewpoint was also used to analyze the Quantile version of Trend Filtering as in Padilla and Chatterjee (2020) and we borrow some of the techniques from Padilla and Chatterjee (2020). The mean regression version of Dyadic CART was throroughly analyzed in Chatterjee and Goswami (2019). We also use some proof techniques developed there and adapt it to our setting. We now discuss a sketch of the proof of our main result in Theorem 1. We have divided the proof sketch into several steps for the convenience of the reader.

From the M estimation viewpoint, the natural loss function which arises is the following population quantile loss function $M : \mathbb{R}^{L_{d,n}} \to \mathbb{R}$

$$M(\theta) := \sum_{i \in L_{d,n}} \mathbb{E}\left(\rho_\tau(y_i - \theta_i) - \rho_\tau(y_i - \theta_i^*)\right).$$

Another loss function that plays a role in our proof is the following Huber loss type function

$$\Delta^2(\theta) = \sum_{i \in L_{d,n}} \min\{|\theta_i|, |\theta_i|^2\}.$$

The Huber loss function Δ^2 is always upper bounded by the population quantile loss; this is the content of Lemma 13 in Padilla and Chatterjee (2020)

which says that, under Assumption 1, there exists an absolute constant $c_0 > 0$ such that for all $\delta \in \mathbb{R}^{L_{d,n}}$ it holds that

$$M(\theta^* + \delta) \geq c_0 \Delta^2(\delta). \tag{B.1}$$

With the notation above in hand, we now proceed to sketch the different steps involved in the proof of Theorem 1.

Step 1: (Preliminary Localization) We first show that $\|\hat{\theta}_{rdp}\|_\infty \leq U$ with high probability. The reader can think of this as a preliminary localization step.

Recall U from Definition 1. Within this proof sketch, we will assume $U = O(1)$ which is the regime of interest. For sequences a_n and b_n, we will also use the notation $a_n \lesssim b_n$ to denote that $a_n \leq C b_n$ for an absolute constant C.

This preliminary localization step is crucial in our proof because it allows us to conclude that

$$\|\hat{\theta}_{rdp} - \theta^*\|^2 \lesssim \Delta^2(\hat{\theta}_{rdp} - \theta^*) \lesssim M(\hat{\theta}_{rdp}) \lesssim \|\hat{\theta}_{rdp} - \theta^*\|^2. \tag{B.2}$$

The last inequality above follows from Lemma 8. The above means that the loss functions M, Δ^2 are essentially equivalent (up to constants) to the squared loss. In the reminder of the proof sketch all the events are intersected with $\|\hat{\theta}_{rdp}\|_\infty \leq U$ as the complement event $\|\hat{\theta}_{rdp}\|_\infty > U$ has negligible probability and can be handled separately.

Step 2: (Reduction to Bounding M loss) Because of (B.2), in order to bound $\mathbb{P}(\|\hat{\theta}_{rdp} - \theta^*\|^2 > t^2)$ for any $t > 0$, it suffices for us to bound

$$\mathbb{P}(M(\hat{\theta}_{rdp}) > t^2).$$

Step 3: (Peeling) To bound the required probability, we perform the peeling step which is a standard step in empirical process theory. We write

$$\mathbb{P}(M(\hat{\theta}_{rdp}) > t^2) = \sum_{j=1}^{J} p_j, \quad p_j := \mathbb{P}(2^{j-1}t^2 < M(\hat{\theta}_{rdp}) \leq 2^j t^2).$$

Again, because $\|\hat{\theta}_{rdp}\|_\infty \leq U$ is true it follows that $M(\hat{\theta}) \lesssim \Delta^2(\hat{\theta}_{rdp} - \theta^*) \lesssim N$ as explained in (C.12). Therefore, we only need to sum up to $J = O(\log N) = \tilde{O}(1)$ in our peeling step.

Step 4: (Basic Inequality, Suprema and Markov's Inequality) For $i \in L_{d,n}$, define the sample version of the quantile population loss function as a random function $\hat{M} : \mathbb{R} \to \mathbb{R}$ such that

$$\hat{M}(\theta) := \sum_{i \in L_{d,n}} \left\{ \rho_\tau(y_i - \theta_i) - \rho_\tau(y_i - \theta_i^*) \right\}.$$

Now the so called basic inequality gives us

$$\hat{M}(\tilde{\theta}) - \hat{M}(\hat{\theta}_{rdp}) + \lambda k_{rdp}(\tilde{\theta}) - \lambda k_{rdp}(\hat{\theta}_{rdp}) \geq 0 \tag{B.3}$$

for any $\tilde{\theta} \in \Theta$, with Θ the parameter space defined in (C.3).

Now we can bound p_j as follows. For any reference $\tilde{\theta} \in \Theta$, notice that

$$p_j = \mathbb{P}(2^{j-1}t^2 < M(\hat{\theta}_{rdp}), M(\hat{\theta}_{rdp}) \leq 2^j t^2,) \leq$$
$$\mathbb{P}(2^{j-1}t^2 < M(\hat{\theta}_{rdp}) + \underbrace{\hat{M}(\tilde{\theta}) - \hat{M}(\hat{\theta}_{rdp}) + \lambda k_{rdp}(\tilde{\theta}) - \lambda k_{rdp}(\hat{\theta}_{rdp})}\underbrace{-M(\tilde{\theta}) + M(\tilde{\theta})},$$
$$M(\hat{\theta}_{rdp}) \leq 2^j t^2)$$

Above, we used the basic inequality and added and subtracted $M(\tilde{\theta})$ because we would like to obtain an oracle risk bound with respect to $\tilde{\theta}$. Now we will just "*sup out*" $\hat{\theta}_{rdp}$ to obtain

$$p_j \leq \mathbb{P}\Bigg(2^{j-1}t^2 < \sup_{\theta\in\Theta:\|\theta-\theta^*\|^2\lesssim 2^j t^2} \big(M(\theta) + \hat{M}(\tilde{\theta}) - \hat{M}(\theta) + \lambda k_{rdp}(\tilde{\theta}) - \lambda k_{rdp}(\theta) - M(\tilde{\theta}) + M(\tilde{\theta})\big)\Bigg)$$

where again we used the equivalence of the squared loss and the M loss. Next, we simply use Markov's inequality to obtain $p_j \leq T_{1,j} + T_2$, where

$$T_{1,j} := \tfrac{1}{2^{j-1}t^2}\mathbb{E}\Bigg(\sup_{\theta\in\Theta:\|\theta-\theta^*\|^2\lesssim 2^j t^2} \big(M(\theta) - \hat{M}(\theta) - M(\tilde{\theta}) + \hat{M}(\tilde{\theta}) + \lambda k_{rdp}(\tilde{\theta}) - \lambda k_{rdp}(\theta)\big)\Bigg),$$

and

$$T_{2,j} = \frac{M(\tilde{\theta})}{2^{j-1}t^2}.$$

Step 5: (Symmetrization and Contraction) At this point, $T_{1,j}$ can be viewed as the expected suprema of a penalized empirical process. To simplify matters further, we use the tools of symmetrization (Lemma 3) and contraction (Lemma 4) to convert a penalized empirical process to a penalized Rademacher process. Specifically, we show that

$$T_{1,j} \lesssim T'_{1,j} := \frac{1}{2^{j-1}t^2}\mathbb{E}\left(\sup_{\theta\in\Theta\,:\,\|\theta-\theta^*\|^2\lesssim 2^j t^2}\left\{\xi^\top(\theta-\tilde{\theta}) + \frac{\lambda}{2}(k_{rdp}(\tilde{\theta}) - k_{rdp}(\theta))\right\}\right) \tag{B.4}$$

where ξ is a vector of independent Radamacher random variables, see Lemma 7.

Step 6: (Bounding Penalized Rademacher Complexity) At this point, we are left with the task of bounding the penalized Rademacher complexity term as in (B.4). This task has essentially been done in Proposition 8.9 in Chatterjee and Goswami (2019), the only difference being that they bounded the corresponding Gaussian complexity term. It is not hard to convert the ideas there to our setting where we have Rademacher variables. In lemma 7 we show that

$$T'_{1,j} = \tilde{O}\left(\frac{k_{rdp}(\tilde{\theta}) + \|\tilde{\theta} - \theta^*\|^2}{2^{j-1}t^2}\right),$$

provided that λ is chosen to be not less than $O(\log^2 N)$.

Step 7: (Putting everything together) Next, we bound $T_{2,j}$ by the squared error loss using Lemma 8, which shows that

$$M(\tilde{\theta}) \lesssim \|\tilde{\theta} - \theta^*\|^2.$$

It follows from this and the previous steps that

$$\mathbb{P}(\|\hat{\theta}_{rdp} - \theta^*\|^2 > t^2) \leq \sum_{j=1}^{J} p_j \lesssim \sum_{j=1}^{J} \frac{1}{2^{j-1}} \left[\frac{k_{rdp}(\tilde{\theta})}{t^2} + \frac{1}{t^2}\|\tilde{\theta} - \theta^*\|^2 \right],$$

and so we can take an infimum over $\tilde{\theta} \in \Theta$ on the right hand side above. A simple approximation lemma (see Lemma 9) is now used to justify that we can actually take an infimum over all $\tilde{\theta} \in \mathbb{R}^{L_{d,n}}$ in the previous display which leads to

$$\mathbb{P}(\|\hat{\theta}_{rdp} - \theta^*\|^2 > t^2) \lesssim \frac{Q(\theta^*)}{t^2}.$$

where

$$Q_{rdp}(\theta^*) \lesssim \inf_{\theta \in \mathbb{R}^{L_{d,n}}} \left(k_{rdp}(\tilde{\theta}) + \|\tilde{\theta} - \theta^*\|^2 \right).$$

The above display implies that $\|\hat{\theta}_{rdp} - \theta^*\|^2 = O_{\mathbb{P}}(Q_{rdp}(\theta^*))$. This concludes the proof.

Appendix C: Proofs

C.1. General estimator

Let $\mathcal{S}$ be a collection of linear subspaces of $\mathbb{R}^N$. For any subspace $S \in \mathcal{S}$ we denote its dimension with $\mathrm{Dim}(S)$ and define a penalty function $k_{\mathcal{S}} : \mathbb{R}^{L_{d,n}} \to \mathbb{Z}_+$ induced by $\mathcal{S}$ as

$$k_{\mathcal{S}}(\theta) := \min\{\mathrm{Dim}(S) : \theta \in S,\ S \in \mathcal{S}\}, \tag{C.1}$$

with the convention that the minimum of the empty set is ∞. We are interested in subspaces of arrays which are piecewise constant on a rectangular partition of $L_{d,n}$. We denote by Π_S the rectangular partition of $L_{d,n}$ so that S is the subspace of arrays which are constant on every rectangle of Π_S. We will denote a generic rectangle of $L_{d,n}$ by R. When we say $R \in \Pi_S$ we are referring to a rectangle R of the partition Π_S.

The collection of partitions $\mathcal{P}_{rdp}$ or $\mathcal{P}_{\mathrm{hier}}$ will give rise to corresponding collections of linear subspaces $\mathcal{S}$ and the associated complexity measures $k_{rdp}(\theta)$ and $k_{tree}(\theta)$ respectively. For a collection of subspaces $\mathcal{S}$ corresponding to a collection of rectangular partitions, we now define an additional function $s_{\mathcal{S}} : \mathbb{R}^{L_{d,n}} \to \mathbb{Z}^+$ as follows:

Definition 2.

$$s_{\mathcal{S}}(\theta) := \min_{R\,:\,R\in\Pi_S,\ \theta\in S, S\in\mathcal{S},\ \text{and } k_{\mathcal{S}}(\theta)=|\Pi_S|} |R|.$$

In words, $s_{\mathcal{S}}(\theta)$ is the size of the minimal rectangle of the minimal rectangular partition Π_S within $S \in \mathcal{S}$ such that θ is constant on every rectangle of Π_S.

For a given collection of subspaces $\mathcal{S}$ corresponding to a collection of rectangular partitions of $L_{d,n}$ and a constant $c_1 > 0$, we will consider the general $0 < \tau < 1$ quantile estimator

$$\hat{\theta}^{(\tau)}_{\mathcal{S},c_1} = \operatorname*{arg\,min}_{\theta\in\Theta} \left\{ \sum_{i\in L_{d,n}} \rho_\tau(y_i - \theta_i) + \lambda k_{\mathcal{S}}(\theta) \right\}, \tag{C.2}$$

where $\lambda > 0$ is a tuning parameter and

$$\Theta = (\cup_{S\in\mathcal{S}} S) \cap \left\{ \theta \in \mathbb{R}^{L_{d,n}} \,:\, s_{\mathcal{S}}(\theta) \geq c_1 \log N \right\}, \tag{C.3}$$

C.2. Preliminary lemmas

Throughout the proof, without loss of generality we assume that $\tau \leq 0.5$. The case $\tau > 0.5$ can be handled similarly. Throughout we will suppose that Assumption 1 holds. We will also drop the subscript in $\hat{\theta}_{\mathcal{S},c_1}$ and just write $\hat{\theta}$ to avoid notational clutter.

Definition 3. *Let $b > 1$ be fixed and $\theta', \tilde{\theta} \in \mathbb{R}^{L_{d,n}}$ be vectors of τ/b-quantiles and $(1-\tau)/b$-quantiles of the data vector y respectively, so that*

$$\theta'_i = F^{-1}_{y_i}(\tau/b)$$

and

$$\theta''_i = F^{-1}_{y_i}((1-\tau)/b).$$

Then we denote

$$U := \max\{\|\theta'\|_\infty, \|\theta''\|_\infty\}.$$

Remark 9. *The sequence U is clearly a function of set of marginal distributions of y. For realistic distributions the sequence $U = O(1)$. For instance, if we assume that the error distribution is i.i.d then U is a constant sequence.*

Lemma 1. *For any $\alpha > 0$, if we set $c_1 \geq (\alpha+2)b^2/(2(b-1)^2\tau^2)$ then we have*

$$\mathbb{P}(\|\hat{\theta}\|_\infty \leq U) \geq 1 - 2N^{-\alpha}. \tag{C.4}$$

Furthermore, if $y_i - \theta^*_i$ is subGaussian(v) for all $i \in L_{d,n}$ and some $v > 0$, then, the constraint $|R| \geq \gamma$ can be removed from the estimator and it holds that

$$\mathbb{P}\left(\|\hat{\theta}\|_\infty \leq U'\right) \geq 1 - N^{-\alpha}$$

for any $\alpha > 0$, where $U' = U + v\sqrt{2(\alpha+1)\log N}$.

Proof. We first prove the general case in (C.4). Let $\hat{\Pi}$ be the optimal partition on which $\hat{\theta}$ is piecewise constant. Fix a $v \in L_{d,n}$ and denote by R the rectangle of the partition $\hat{\Pi}$ containing v. We know that $\hat{\theta}_v$ is a sample τ quantile of the observations in y_R. Therefore, we can assert that

$$\begin{aligned}&\left\{\exists v \in L_{d,n} \,:\, \hat{\theta}_v < \min_{u \in L_{d,n}} \theta'_u\right\}\\ &\subset \bigcup_{R \subset L_{d,n} \text{ rectangle } :|R| \geq c_1 \log N} \left\{\left|\left\{i \in R \,:\, y_i \leq \min_{u \in L_{d,n}} \theta'_u\right\}\right| > \tau |R|\right\}.\end{aligned}$$

Now for any rectangle R with $|R| \geq c_1 \log N$,

$$\begin{aligned}\mathbb{P}\left(\left|\left\{i \in R \,:\, y_i \leq \min_{j=1,\ldots,n} \theta'_j\right\}\right| > \tau |R|\right) &\leq \mathbb{P}\left(\sum_{u \in R} 1(y_u \leq \theta'_u) > \tau |R|\right)\\ &\leq \exp\left(-\frac{2(b-1)^2}{b^2}\tau^2 |R|\right)\\ &\leq \exp\left(-\tfrac{2(b-1)^2}{b^2}\tau^2 c_1 \log N\right),\end{aligned}$$

where the second inequality follows by Hoeffding's inequality. By choosing $c_1 = (\alpha+2)b^2/(2(b-1)^2\tau^2)$ and by using the last two displays imply that

$$\mathbb{P}(\exists v \,:\, \hat{\theta}_v < \min_{u \in L_{d,n}} \theta'_u) \leq N^{-\alpha}.$$

The same bound can readily be shown for $\mathbb{P}(\exists v \,:\, \hat{\theta}_v > \max_{u \in L_{d,n}} \theta''_u)$ by a similar argument. Both these assertions with a further union bound finishes the proof for the general case.

Let us now assume that $y_i - \theta^*_i$ is subGaussian(v) for all $i \in L_{d,n}$ and some $v > 0$ Then, by construction of $\hat{\theta}$, it holds that

$$\|\hat{\theta}\|_\infty \leq \|y\|_\infty \leq \|\theta^*\|_\infty + \|y - \theta^*\|_\infty \leq U + \|y - \theta^*\|_\infty$$

and by the subGaussian maximal inequality (e.g. see Theorem 1.14 in Rigollet and Hütter (2015)) we have

$$\mathbb{P}\left(\|y - \theta^*\|_\infty \geq v\sqrt{2(\alpha+1)\log N}\right) \leq \frac{1}{N^\alpha}$$

so the proof follows. □

Our next result is a modified version of Lemma 9.1 in Chatterjee and Goswami (2019) where we prove the corresponding result for rademacher random variables instead of gaussian random variables.

Lemma 2 (Lemma 9.1 in Chatterjee and Goswami (2019))**.** *Let $\tilde{\theta} \in \mathbb{R}^{L_{d,n}}$ be a fixed array and $\xi \in \mathbb{R}^{L_{d,n}}$ be an array of independent Rademacher random variables. Then there exists $C > 0$ such that if $\lambda \geq C \log N$ then it follows that*

$$\mathbb{E}\left(\sup_{\theta \in \Theta}\left\{\left(\xi^\top \frac{(\theta - \tilde{\theta})}{\|\theta - \tilde{\theta}\|}\right)^2 - \lambda k_{\mathcal{S}}(\theta)\right\}\right) \leq 16.$$

Proof. Let $S \in \mathcal{S}$ and O_S be the orthogonal projection matrix onto S. Then, following the arguments in the proof of Lemma 9.1 in Chatterjee and Goswami (2019) it follows that

$$\sup_{v\neq 0, v\in S} \frac{\xi^\top (v-\tilde{\theta})}{\|v-\tilde{\theta}\|} \leq \frac{\xi^\top (I-O_S)\tilde{\theta}}{\|(I-O_S)\tilde{\theta}\|} + \sup_{v\in S, \|v\|\leq 1} \xi^\top v. \tag{C.5}$$

Let $v_{S,1}, \ldots, v_{S,m_S}$ an orthonormal basis of S with $m_S = \dim(S)$. Then

$$\begin{aligned}
\sup_{v\in S, \|v\|\leq 1} (\xi^\top v)^2 &= \sup_{v\in S, \|v\|\leq 1} (\xi^\top O_S v)^2 \\
&\leq \sup_{v\in S, \|v\|\leq 1} \|O_S\xi\|^2 \cdot \|v\|^2 \\
&= \|O_S\xi\|^2 \\
&= \left\| \sum_{j=1}^{m_S} (v_{S,j}^\top \xi) v_j \right\|^2 \\
&= \sum_{j=1}^{m_s} |v_{S,j}^\top \xi|^2 \\
&\leq m_S \max_{j=1,\ldots,m_S} |Z_j^{(S)}|^2
\end{aligned}$$

where $Z_j^{(S)} := v_{S,j}^\top \xi$ is Sub-Gaussian(1). Therefore,

$$\begin{aligned}
\mathbb{E}\left(\sup_{v\in S, \|v\|\leq 1} \xi^\top v \right) &\leq \sqrt{\dim(S)} \cdot \mathbb{E}\left(\max_{j=1,\ldots,m_S} |Z_j^{(S)}| \right) \\
&\leq \sqrt{2\dim(S)\log N},
\end{aligned} \tag{C.6}$$

where the second inequality holds by the usual expectation of maximum of sub-Gaussian random variables inequality. Then from an application of McDiarmid's inequality inequality as in Page 62 of van Handel (2014), we obtain that for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{v\in S, \|v\|\leq 1} (\xi^\top v)^2 - 4\dim(S)\log N \geq 2t \right) \\
&\leq \mathbb{P}\left(\sup_{v\in S, \|v\|\leq 1} (\xi^\top v) - \sqrt{2\dim(S)\log N} \geq \sqrt{t} \right) \\
&\leq \exp\left(-\frac{t}{4}\right).
\end{aligned}$$

Hence, by union bound and the fact that $|\{S \in \mathcal{S}, \dim(S) = k\}| \leq N^{2k}$, we have that

$$\begin{aligned}
&\mathbb{P}\left(\max_{S\in\mathcal{S}, \dim(S)=k} \left\{ \sup_{v\in S, \|v\|\leq 1} (\xi^\top v)^2 - 20\dim(S)\log N \right\} \geq 2t \right) \\
&\leq \exp\left(-\frac{t}{4}\right).
\end{aligned}$$

And so again, by union bound we obtain that

$$\mathbb{P}\left(\max_{k=1,\ldots,N}\ \max_{S\in\mathcal{S},\dim(S)=k}\left\{\sup_{v\in S,\|v\|\le 1}(\xi^\top v)^2-28\dim(S)\log N\right\}\ge 2t\right)\le\exp\left(-\frac{t}{4}\right). \tag{C.7}$$

Similarly,

$$\mathbb{P}\left(\max_{k=1,\ldots,N}\ \max_{S\in\mathcal{S},\dim(S)=k}\left\{\left(\frac{\xi^\top(I-O_S)\tilde{\theta}}{\|(I-O_S)\tilde{\theta}\|}\right)^2-28\log N\right\}\ge 2t\right)\le\exp\left(-\frac{t}{4}\right). \tag{C.8}$$

The claim follows from (C.7) and (C.8) by simple integration. □

Next, we recall some notations from Section B.

Definition 4. *For $u\in L_{d,n}$, define the random function $\hat{M}_u:\mathbb{R}\to\mathbb{R}$ as follows:*

$$\hat{M}_u(\theta_u):=\{\rho_\tau(y_u-\theta_u)-\rho_\tau(y_u-\theta_u^*)\},$$

Now define the random function $\hat{M}:\mathbb{R}\to\mathbb{R}$

$$\hat{M}(\theta):=\sum_{u\in L_{d,n}}\hat{M}_u(\theta_u),$$

and the deterministic function $M:\mathbb{R}\to\mathbb{R}$ as

$$M(\theta):=\mathbb{E}\left(\hat{M}(\theta)\right).$$

Furthermore, let us denote

$$\Delta^2(\theta)=\sum_{i\in L_{d,n}}\min\{|\theta_i|,|\theta_i|^2\}$$

and $\Delta_N^2(\theta)=\Delta^2(\theta)/N$.

Our analysis relies on viewing the estimator defined in (C.2) as a penalized M estimator or a penalized empirical risk minimization estimator. Hence the natural loss function for us is the population quantile loss M function given above. However, we would like to give risk bounds for the square loss function. For this purpose, the Δ function defined above plays an important role in converting bounds in the M loss function to bounds for squared error loss.

We now proceed to state some results (Lemmas 8–11) involving involving the functions $M(\cdot)$ and $\Delta^2(\cdot)$. These are results that also appeared in Padilla and Chatterjee (2020), the only difference with the results in Padilla and Chatterjee (2020) is that we now use the penalty function $k_{\mathcal{S}}(\theta)$ instead of the TV(θ) function. We omit writing the proofs of these results since the proofs are very similar to what is already given in Padilla and Chatterjee (2020).

Lemma 3. *(Symmetrization, Lemma 28 in Padilla and Chatterjee (2020)). For any set K, $\tilde{\theta} \in \mathbb{R}^{L_{d,N}}$, and $\lambda > 0$ it holds that*

$$E\left[\sup_{\theta\in K}\left\{M(\theta) - M(\tilde{\theta}) + \hat{M}(\tilde{\theta}) - \hat{M}(\theta) + \lambda(k_{\mathcal{S}}(\tilde{\theta}) - k_{\mathcal{S}}(\theta))\right\}\right]$$
$$\leq 2\,E\left[\sup_{\theta\in K}\left\{\sum_{i\in L_{d,n}} \xi_i(\hat{M}_i(\theta_i) - \hat{M}_i(\tilde{\theta}_i)) + \frac{\lambda}{2}k_{\mathcal{S}}(\tilde{\theta}) - \frac{\lambda}{2}k_{\mathcal{S}}(\theta)\right\}\right],$$

where $\xi_1, \ldots, \xi_n$ are independent Rademacher variables independent of $\{y_i\}_{i=1}^n$.

Lemma 4. *(Contraction principle, Lemma 29 in Padilla and Chatterjee (2020)). Let $h_1, \ldots, h_n : \mathbb{R} \to \mathbb{R}$ η-Lipschitz functions for some $\eta > 0$. Then for any $\tilde{\theta} \in \mathbb{R}^{L_{d,n}}$, any compact set K, and $\xi_1, \ldots, \xi_n$ independent Rademacher variables we have that*

$$E\left[\sup_{\theta\in K}\left\{\sum_{i\in L_{d,n}} \xi_i h_i(\theta_i) \; + \frac{\lambda}{2}k_{\mathcal{S}}(\tilde{\theta}) - \frac{\lambda}{2}k_{\mathcal{S}}(\theta)\right\}\right]$$
$$\leq\; E\left[\sup_{\theta\in K}\left\{\eta\sum_{i\in L_{d,n}} \xi_i\theta_i \; + \frac{\lambda}{2}k_{\mathcal{S}}(\tilde{\theta}) - \frac{\lambda}{2}k_{\mathcal{S}}(\theta)\right\}\right]$$

for any $\lambda > 0$.

Lemma 5. *(Lemma 13 in Padilla and Chatterjee (2020)). Suppose that Assumption 1 holds. Then there exists a constant c_0 such that for all $\delta \in R^n$, we have*

$$M(\theta^* + \delta) \geq c_0\Delta^2(\delta).$$

Lemma 6. *(Lemma 17 in Padilla and Chatterjee (2020)). Let $\delta \in R^n$. Then*

$$\|\delta\|^2 \;\leq\; \max\{\|\delta\|_\infty, 1\}\Delta^2(\delta). \tag{C.9}$$

Our next lemma is key and controls the expected suprema of a penalized empirical process.

Lemma 7. *Let $\tilde{\theta} \in \mathbb{R}^n$ and $t > 0$. Then there exist a constant $C > 0$ such that for any $a > 0$ if $\lambda \geq Ca\log N$, we have that*

$$\mathbb{E}\left(\sup_{\theta\in\Theta\,:\,\|\theta-\theta^*\|^2\leq t^2}\left\{M(\theta) - M(\tilde{\theta}) + \hat{M}(\tilde{\theta}) - \hat{M}(\theta) + \lambda(k_{\mathcal{S}}(\tilde{\theta}) - k_{\mathcal{S}}(\theta))\right\}\right)$$
$$\leq\; C_2 a + \frac{2t^2}{a} + \frac{2\|\theta^* - \tilde{\theta}\|^2}{a} + \lambda k_{\mathcal{S}}(\tilde{\theta}),$$

for a positive constant $C_2 > 0$.

Proof. Notice that if $\xi \in \mathbb{R}^{L_{d,n}}$ consists of independent Rademacher random variables then

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\theta\in\Theta\,:\,\|\theta-\theta^*\|\le t}\left\{M(\theta)-M(\tilde\theta)+\hat M(\tilde\theta)-\hat M(\theta)+\lambda(k_{\mathcal S}(\tilde\theta)-k_{\mathcal S}(\theta))\right\}\right)\\
&\le 2\,\mathbb{E}\left(\sup_{\theta\in\Theta\,:\,\|\theta-\theta^*\|\le t}\left\{\xi^\top(\theta-\tilde\theta)+\frac{\lambda}{2}(k_{\mathcal S}(\tilde\theta)-k_{\mathcal S}(\theta))\right\}\right)\\
&\le 2\,\mathbb{E}\left(\sup_{\theta\in\Theta\,:\,\|\theta-\theta^*\|\le t}\left\{\xi^\top(\theta-\tilde\theta)-\frac{\lambda}{2}k_{\mathcal S}(\theta)\right\}\right)+\lambda k_{\mathcal S}(\tilde\theta)\\
&\le 2\,\mathbb{E}\left(\sup_{\theta\in\Theta\,:\,\|\theta-\theta^*\|\le t}\left\{\frac{a}{2}\left(\xi^\top\frac{(\theta-\tilde\theta)}{\|\theta-\tilde\theta\|}\right)^2-\frac{\lambda}{2}k_{\mathcal S}(\theta)\right\}\right)\\
&\quad+\frac{2t^2+2\|\theta^*-\tilde\theta\|^2}{a}+\lambda k_{\mathcal S}(\tilde\theta)\\
&\le a\,\mathbb{E}\left(\sup_{\theta\in\Theta\,:\,\|\theta-\theta^*\|\le t}\left\{\left(\xi^\top\frac{(\theta-\tilde\theta)}{\|\theta-\tilde\theta\|}\right)^2-Ck_{\mathcal S}(\theta)\log N\right\}\right)+\frac{2t^2}{a}+\frac{2\|\theta^*-\tilde\theta\|^2}{a}\\
&\quad+\lambda k_{\mathcal S}(\tilde\theta)\\
&\le C_2a+\frac{2t^2}{a}+\frac{2\|\theta^*-\tilde\theta\|^2}{a}+\lambda k_{\mathcal S}(\tilde\theta)
\end{aligned}$$

where the first inequality follows as in Lemmas 3 and 4, the third by Cauchy Schwarz inequality, and the last by Lemma 2. □

Lemma 8. *For $\tilde\theta\in\mathbb{R}^N$ we have that*

$$M(\tilde\theta)\le\frac{\overline f}{2}\|\tilde\theta-\theta^*\|^2.$$

Proof. Let $\delta:=\tilde\theta-\theta^*$. We start by recalling by Equation (19) in Padilla and Chatterjee (2020) which states that

$$M_i(\tilde\theta_i)=\int_0^{\delta_i}(F_{y_i}(\theta_i^*+z)-F_{y_i}(\theta_i^*))dz.$$

Hence,

$$\begin{aligned}
M(\tilde\theta) &= \sum_{i\in L_{d,n}}\int_0^{\delta_i}(F_{y_i}(\theta^*+z)-F_{y_i}(\theta^*))\,dz\\
&\le \sum_{i\in L_{d,n}}\int_0^{\delta_i}\overline f z dz\\
&= \frac{\overline f}{2}\|\delta\|^2
\end{aligned}$$

where the inequality follows from Assumption 1. □

C.3. General upper bound

Theorem 4. *Suppose that Assumption 1 holds. There exists universal constants $c_1, C_1, C_2, C_3 > 0$ such that for any $0 < \epsilon < 1$, if we set $\gamma = c_1 \log N$ and*

$$\lambda \;=\; C_1 \frac{\max\{1, U\} \log(N) \log(NU)}{\epsilon},$$

implies that with probability at least $1 - C_2\epsilon$,

$$\frac{\|\hat{\theta} - \theta^*\|^2}{N} \leq \frac{C_3 Q(\theta^*)}{\epsilon^2}, \tag{C.10}$$

where

$$Q(\theta^*) \;:=\; \inf_{\theta \in \Theta} \left\{ \frac{k_{\mathcal{S}}(\theta) \max\{1, U^2\} \log^2(\max\{N, U\})}{N} \;+ \frac{\overline{f} \|\theta^* - \theta\|^2}{N} \right\}.$$

Proof. Let $t \in (0, 2N\,U)$ and notice that for $\hat{\delta} := \hat{\theta} - \theta^*$ we have that for U as in Definition 1 it holds that

$$\mathbb{P}(\Delta^2(\hat{\delta}) > t^2) \;\leq\; \mathbb{P}(\Delta^2(\hat{\delta}) > t^2, \|\hat{\theta}\|_\infty \leq U) + \mathbb{P}(\|\hat{\theta}\|_\infty > U), \tag{C.11}$$

with $\Delta(\cdot)$ as in Definition 3. Hence, from Lemma 1 it is enough to bound $\mathbb{P}(\Delta^2(\hat{\delta}) > t^2, \|\hat{\theta}\|_\infty \leq U)$. Towards that end we notice that $\|\hat{\theta}\|_\infty \leq U$ implies

$$\Delta^2(\hat{\delta}) \;\leq\; \|\hat{\delta}\|_1 \;\leq\; \|\hat{\theta}\|_1 + \|\theta^*\|_1 \leq 4N\,U \tag{C.12}$$

and hence

$$\mathbb{P}(\Delta^2(\hat{\delta}) > t^2, \|\hat{\theta}\|_\infty \leq U) \quad\leq\quad \mathbb{P}(\Delta^2(\hat{\delta}) > t^2, \Delta^2(\hat{\delta}) \leq 4N\,U, \|\hat{\theta}\|_\infty \leq U).$$

Now, we will undertake the so called peeling step. Letting

$$x \;=\; \lceil \log(4N\,U/t^2)/\log 2 \rceil, \tag{C.13}$$

we obtain that for any $\tilde{\theta} \in \Theta$

$$\begin{aligned}
&\mathbb{P}(\Delta^2(\hat{\delta}) > t^2, \|\hat{\theta}\|_\infty \leq U) \\
&\leq \sum_{j=1}^{x} \mathbb{P}\left(\Delta^2(\hat{\delta}) > 2^{j-1}t^2, \Delta^2(\hat{\delta}) \leq 2^j t^2, \|\hat{\theta}\|_\infty \leq U\right) \\
&\leq \sum_{j=1}^{x} \mathbb{P}\left(M(\hat{\theta}) > c_0 2^{j-1}t^2, \Delta^2(\hat{\delta}) \leq 2^j t^2, \|\hat{\theta}\|_\infty \leq U\right) \\
&= \sum_{j=1}^{x} \mathbb{P}(M(\hat{\theta}) - M(\tilde{\theta}) + M(\tilde{\theta}) > c_0 2^{j-1}t^2, \Delta^2(\hat{\delta}) \leq 2^j t^2, \\
&\|\hat{\theta}\|_\infty \leq U) \\
&\leq \sum_{j=1}^{x} \mathbb{P}(M(\hat{\theta}) - M(\tilde{\theta}) + \\
&\left\{\hat{M}(\tilde{\theta}) - \hat{M}(\hat{\theta}) + \lambda(k_{\mathcal{S}}(\tilde{\theta}) - k_{\mathcal{S}}(\hat{\theta}))\right\} + M(\tilde{\theta}) \\
&> c_0 2^{j-1}t^2, \Delta^2(\hat{\delta}) \leq 2^j t^2, \|\hat{\theta}\|_\infty \leq U)
\end{aligned}$$

where the second inequality follows from Lemma 5, and the last by the optimality of $\hat{\theta}$, since

$$\hat{M}(\tilde{\theta}) - \hat{M}(\hat{\theta}) + k_{\mathcal{S}}(\tilde{\theta}) - k_{\mathcal{S}}(\hat{\theta}) \geq 0.$$

We can now continue to write from the previous display,

$$\begin{aligned}
&\mathbb{P}(\Delta^2(\hat{\delta}) > t^2, \|\hat{\theta}\|_\infty \leq U) \\
&\leq \sum_{j=1}^{x} \mathbb{P}\Bigg(\sup_{\theta \in \Theta : \Delta^2(\theta - \theta^*) \leq 2^j t^2, \|\theta\|_\infty \leq U} \Big\{ M(\theta) - M(\tilde{\theta}) + \hat{M}(\tilde{\theta}) - \hat{M}(\theta) + \\
&\qquad \lambda(k_{\mathcal{S}}(\tilde{\theta}) - k_{\mathcal{S}}(\theta)) + M(\tilde{\theta}) \Big\} \geq c_0 2^{j-1} t^2 \Bigg) \\
&\leq \sum_{j=1}^{x} \mathbb{P}\Bigg(\sup_{\theta \in \Theta : \|\theta - \theta^*\|^2 \leq 2^j \max\{1,U\} t^2} \Big\{ M(\theta) - M(\tilde{\theta}) + \hat{M}(\tilde{\theta}) - \hat{M}(\theta) + \\
&\qquad \lambda(k_{\mathcal{S}}(\tilde{\theta}) - k_{\mathcal{S}}(\theta)) + M(\tilde{\theta}) \Big\} \geq c_0 2^{j-1} t^2 \Bigg) \\
&\leq \sum_{j=1}^{x} \frac{1}{c_0 2^{j-1} t^2} \mathbb{E}\Bigg(\sup_{\theta \in \Theta : \|\theta - \theta^*\|^2 \leq 2^j \max\{1,U\} t^2} \Big\{ M(\theta) - M(\tilde{\theta}) + \\
&\qquad \hat{M}(\tilde{\theta}) - \hat{M}(\theta) + \lambda(k_{\mathcal{S}}(\tilde{\theta}) - k_{\mathcal{S}}(\theta)) \Big\} \Bigg) + \frac{2M(\tilde{\theta})}{c_0 t^2}
\end{aligned} \tag{C.14}$$

where the second inequality follows from Lemma 6, and the third inequality follows from Markov's inequality and summing up the geometric series.

Let $\epsilon \in (0,1)$ be fixed. We notice that (C.14) and Lemma 8 imply that

$$\begin{aligned}
&\mathbb{P}(\Delta^2(\hat{\delta}) > t^2, \|\hat{\theta}\|_\infty \leq U) \\
&\leq \sum_{j=1}^{x} \frac{1}{c_0 2^{j-1} t^2} \mathbb{E}\Bigg(\sup_{\theta \in \Theta : \|\theta - \theta^*\|^2 \leq 2^j \max\{1,U\} t^2} \Big\{ M(\theta) - M(\tilde{\theta}) + \\
&\qquad \hat{M}(\tilde{\theta}) - \hat{M}(\theta) + \lambda(k_{\mathcal{S}}(\tilde{\theta}) - k_{\mathcal{S}}(\theta)) \Big\} \Bigg) + \\
&\qquad \frac{2\overline{f}}{c_0 t^2} \|\tilde{\theta} - \theta^*\|^2.
\end{aligned}$$

Next, for some $a > 0$ to be chosen later, we set $\lambda = Ca \log N$. Hence, from Lemma 7, we have that

$$\begin{aligned}
&\mathbb{P}(\Delta^2(\hat{\delta}) > t^2, \|\hat{\theta}\|_\infty \leq U) \\
&\leq \sum_{j=1}^{x} \frac{1}{c_0 2^{j-1} t^2} \left[C_2 a + \frac{2^{j+1} \max\{1,U\} t^2 + 2\|\tilde{\theta} - \theta^*\|^2}{a} + \lambda k_{\mathcal{S}}(\tilde{\theta}) \right] \\
&+ \frac{2\overline{f}}{c_0 t^2} \|\tilde{\theta} - \theta^*\|^2 \\
&\leq \frac{2C_2 a}{c_0 t^2} + \frac{4 \max\{1,U\} x}{a c_0} + \frac{2\lambda k_{\mathcal{S}}(\tilde{\theta})}{c_0 t^2} + \frac{1}{t^2} \left(\frac{4}{a c_0} + \frac{2\overline{f}}{c_0} \right) \|\tilde{\theta} - \theta^*\|^2.
\end{aligned} \tag{C.15}$$

Therefore, setting

$$a := \frac{\max\{1,U\} \log(NU)}{\epsilon},$$

$$t^2 := \frac{C\max\{1, U, C_2\}k_{\mathcal{S}}(\tilde{\theta})\log^2(NU)}{\epsilon^2 c_0} + \frac{\overline{f}\|\tilde{\theta}-\theta^*\|^2}{2\epsilon},$$

and letting

$$\tilde{\theta} \in \arg\min_{\theta}\left\{\frac{k_{\mathcal{S}}(\theta)\max\{1,U^2\}\log^2(\max\{N,U\})}{N} + \frac{\overline{f}\|\theta^*-\theta\|^2}{N}\right\}$$

we obtain the conclusion in (C.10) by combining (C.11) and (C.15). □

C.4. Proof of Theorem 1

In the rest of the proofs we denote $\hat{\theta}_{rdp}$ simply as $\hat{\theta}$.

Proof. From Theorem 4 we obtain that

$$\frac{\|\hat{\theta}-\theta^*\|^2}{N} = O_{\mathbb{P}}\left(\inf_{\tilde{\theta}\in\Theta}\left\{\frac{\overline{f}\|\tilde{\theta}-\theta^*\|^2}{N} + \frac{k_{rdp}(\tilde{\theta})\max\{1,U^2\}\log^2(\max\{N,U\})}{N}\right\}\right). \tag{C.16}$$

However, for any $\theta \in \mathbb{R}^N$ by Lemma 9 there exists $A(\theta) \in \Theta$ such that $k_{rdp}(A(\theta)) \le k_{rdp}(\theta)$ and

$$\|A(\theta)-\theta\|^2 \le 4c_1\|\theta\|_\infty^2 k_{rdp}(\theta)\log N.$$

It follows that

$$\begin{aligned}
&\inf_{\tilde{\theta}\in\Theta}\left\{\frac{\overline{f}\|\tilde{\theta}-\theta^*\|^2}{N} + \frac{k_{rdp}(\tilde{\theta})\max\{1,U^2\}\log^2(\max\{N,U\})}{N}\right\}\\
&\le \inf_{\theta\in\mathbb{R}^N}\left\{\frac{\overline{f}\|A(\theta)-\theta^*\|^2}{N} + \frac{k_{rdp}(A(\theta))\max\{1,U^2\}\log^2(\max\{N,U\})}{N}\right\}\\
&\le \inf_{\theta\in\mathbb{R}^N}\left\{\frac{2\overline{f}\|A(\theta)-\theta\|^2}{N} + \right.\\
&\qquad \left.\frac{2\overline{f}\|\theta-\theta^*\|^2}{N} + \frac{k_{rdp}(A(\theta))\max\{1,U^2\}\log^2(\max\{N,U\})}{N}\right\}\\
&\le \inf_{\theta\in\mathbb{R}^N}\left\{\frac{8\overline{f}c_1\|\theta\|_\infty^2 k_{rdp}(\theta)\log N}{N} + \frac{2\overline{f}\|\theta-\theta^*\|^2}{N} + \right.\\
&\qquad \left.\frac{k_{rdp}(\theta)\max\{1,U^2\}\log^2(\max\{N,U\})}{N}\right\}
\end{aligned} \tag{C.17}$$

where the second inequality follows from the Cauchy–Schwarz inequality, and the third by the construction of $A(\cdot)$. The claim follows combining (C.16) with (C.17). □

C.5. Other lemmas

Lemma 9. *Let $\theta \in \mathbb{R}^{L_{d,n}}$. Given $c_1 > 0$ there exists a $\tilde{\theta} \in \mathbb{R}^{L_{d,n}}$ such that the following holds:*

- $k_{rdp}(\tilde{\theta}) \leq k_{rdp}(\theta)$.
- $$\|\tilde{\theta} - \theta\|^2 \leq 4c_1 \|\theta\|_\infty^2 k_{rdp}(\theta) \log N.$$
- $s(\tilde{\theta}) \geq c_1 \log N$, *where* $s(\cdot)$ *is the* $s_{\mathcal{S}}(\cdot)$ *corresponding to Dyadic partitions.*

Proof. Let Π a minimal dyadic partition induced by θ. Then consider $\tilde{\Pi}$ the dyadic partition obtained by performing the same splits as in the construction of Π but only when each split produces rectangles of size at least $c_1 \log N$. Then let $\tilde{\theta}$ be constructed by averaging the values of θ on each rectangle of $\tilde{\Pi}$. Notice that by construction the first and third claims of the lemma hold. To see why the second claim holds, we observe that Π and $\tilde{\Pi}$ differ in at most $k_{rdp}(\theta)$ rectangles each of which is of size at most $2c_1 \log N$. The claim then follows. □

C.6. Proof of Corollary 1

Proof. **Case** $d = 1$**.**

We proceed in two cases. First, if $V = \mathrm{TV}(\theta^*) = 0$ then $k_{rdp}(\theta^*) = 1$ and Theorem 1 implies that

$$\frac{1}{N} \sum_{i \in L_{d,N}} (\hat{\theta}_i - \theta_i^*)^2 = O_{\mathbb{P}}\left(\frac{\max\{1, U^2\} \log^2\{N, U\}}{N} \right). \tag{C.18}$$

Suppose now that $V > 0$. Then by Proposition 8.9 in Chatterjee and Goswami (2019), for any $\eta > 0$ there exits θ such that for some positive constant C it holds that $k_{rdp}(\theta) \leq C\eta^{-1}$ and

$$\|\theta - \theta^*\|_\infty \leq V\eta.$$

Then notice that $\|\theta\|_\infty^2 \leq 2V^2\eta^2 + 2U^2$ and $\|\theta - \theta^*\|^2 \leq V^2\eta^2 N$.

Next, we set

$$\eta := \frac{1}{V^{2/3}} \left(\frac{\log N}{N} \right)^{1/3}$$

and notice that

- $$\begin{aligned} \frac{\|\theta\|_\infty^2 k_{rdp}(\theta) \log N}{N} &\leq 2\left(V^2\eta^2 + U^2\right) C\eta^{-1} \log N \\ &\leq \frac{2CV^2\eta \log N}{N} + \frac{2CU^2\eta^{-1} \log N}{N} \\ &= O\left(\frac{U^2 V^{2/3} \log^{2/3} N}{N^{2/3}} \right). \end{aligned} \tag{C.19}$$
- $$\frac{\|\theta - \theta^*\|^2}{N} = V^2\eta^2 = V^2 \left(\frac{1}{V^{2/3}} \left(\frac{\log N}{N} \right)^{1/3} \right)^2 = \frac{V^{2/3} \log^{2/3} N}{N^{2/3}}.$$

- $$\frac{k_{rdp}(\theta)\max\{1,U^2\}\log^2(\max\{N,U\})}{N} \le \frac{CV^{2/3}N^{1/3}(\log^{-1/3}N)\max\{1,U^2\}\log^2\max\{N,U\}}{N} = \frac{CV^{2/3}\max\{1,U^2\}\log^{5/3}\max\{N,U\}}{N^{2/3}}.$$

Combining the cases above we obtain the claim for $d = 1$.

Case $d > 1$**.** If case $V = 0$ we proceed as we did in the previous case. Suppose now that $V > 0$. Then, by the proof of Theorem 4.2 in Chatterjee and Goswami (2019), for any $\eta > 0$ there exists a θ such that

$$k_{rdp}(\theta) \le \frac{C\,\mathrm{TV}(\theta^*)\,\log N}{\eta},$$

$$\|\theta\|_\infty \le \|\theta^*\|_\infty,$$

and

$$\|\theta - \theta^*\|^2 \le C\eta\,\mathrm{TV}(\theta^*)\,\log N$$

for some positive constant C.

Next, let $\eta = \log N$ and notice that

- $$\frac{\|\theta\|_\infty^2 k_{rdp}(\theta)\log N}{N} \le \frac{CV\log N}{\eta}\cdot\frac{U^2\log N}{N} = \frac{CVU^2\log N}{N}. \tag{C.20}$$
- $$\frac{\|\theta-\theta^*\|^2}{N} = \frac{CV\log^2 N}{N}. \tag{C.21}$$
- $$\frac{k_{rdp}(\theta)\max\{1,U^2\}\log^2(\max\{N,U\})}{N} \le \frac{CV\max\{1,U^2\}\log^2(\max\{N,U\})}{N}. \tag{C.22}$$

Therefore, combining (C.20)–(C.22) the claim follows. □

References

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. MR3909963

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011. MR2797841

JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, and Ming-Ming Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4181–4190, 2017.

Gilles Blanchard, Christin Schäfer, Yves Rozenholc, and K-R Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2-3):209–241, 2007.

Halley L Brantley, Joseph Guinness, and Eric C Chi. Baseline drift estimation for air quality data using quantile trend filtering. *arXiv preprint arXiv:1904.10582*, 2019. MR4117821

Lawrence D Brown, T Tony Cai, and Harrison H Zhou. Robust nonparametric estimation via wavelet median regression. *The Annals of Statistics*, 36(5): 2055–2084, 2008. MR2458179

Lorenzo Cappello, Oscar Hernan Madrid Padilla, and Julia A Palacios. Scalable bayesian change point detection with spike and slab priors. *arXiv preprint arXiv:2106.10383*, 2021. MR4669263

Sabyasachi Chatterjee and Subhajit Goswami. Adaptive estimation of multivariate piecewise polynomials and bounded variation functions by optimal decision trees. *arXiv preprint arXiv:1911.11562*, 2019. MR4338374

Sabyasachi Chatterjee and Subhajit Goswami. New risk bounds for 2d total variation denoising. *IEEE Transactions on Information Theory*, 67(6):4060–4091, 2021. MR4289366

Probal Chaudhuri and Wei-Yin Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 8(5):561–576, 2002. MR1935647

Dennis D Cox. Asymptotics for m-type smoothing splines. *The Annals of Statistics*, pages 530–551, 1983. MR0696065

David L Donoho. Cart and best-ortho-basis: a connection. *The Annals of statistics*, 25(5):1870–1911, 1997. MR1474073

Randall L Eubank. *Spline smoothing and nonparametric regression*, volume 90. M. Dekker New York, 1988. MR0934016

Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering. *The Annals of Statistics*, 48(1):205–229, 2020. MR4065159

Xuming He and Peide Shi. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journaltitle of Nonparametric Statistics*, 3(3-4):299–308, 1994. MR1291551

Xuming He, Pin Ng, and Stephen Portnoy. Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60 (3):537–550, 1998. MR1625950

Jan-Christian Hutter and Philippe Rigollet. Optimal rates for total variation denoising. *Annual Conference on Learning Theory*, 29:1115–1146, 2016.

Laura Jula Vanegas, Merle Behr, and Axel Munk. Multiscale quantile segmentation. *Journal of the American Statistical Association*, pages 1–14, 2021. MR4480719

Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994. MR1326417

Andrew Lang, Aaron Carass, Peter A Calabresi, Howard S Ying, and Jerry L Prince. An adaptive grid for graph-based segmentation in retinal oct. In *Medical Imaging 2014: Image Processing*, volume 9034, page 903402. International Society for Optics and Photonics, 2014.

Enno Mammen and Sara van de Geer. Locally apadtive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997. MR1429931

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006. MR2274394

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

Francesco Ortelli and Sara van de Geer. Prediction bounds for higher order total variation regularized least squares. *arXiv preprint arXiv:1904.10871*, 2019. MR4338382

Oscar Hernan Madrid Padilla and Sabyasachi Chatterjee. Risk bounds for quantile trend filtering. *arXiv preprint arXiv:2007.07472*, 2020. MR4472846

Oscar Hernan Madrid Padilla, James Sharpnack, James G. Scott, and Ryan J. Tibshirani. The DFS fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18:176:1–176:36, 2018. MR3827064

Oscar Hernan Madrid Padilla, Wesley Tansey, and Yanzhen Chen. Quantile regression with deep relu networks: Estimators and minimax rates. *arXiv preprint arXiv:2010.08236*, 2020. MR4577686

Oscar Hernan Madrid Padilla, Yi Yu, and Alessandro Rinaldo. Lattice partition recovery with dyadic cart. *arXiv preprint arXiv:2105.13504*, 2021.

Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J. Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. *To appear, Neural Information Processing Systems*, 2016.

Clayton Scott and Robert D Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE transactions on information theory*, 52(4):1335–1353, 2006. MR2241192

Wesley Tansey, Oluwasanmi Koyejo, Russell A Poldrack, and James G Scott. False discovery rate smoothing. *Journal of the American Statistical Association*, 113(523):1156–1171, 2018. MR3862347

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014. MR3189487

Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012. MR2985948

Florencio I Utreras. On computing robust splines and applications. *SIAM Journal on Scientific and Statistical Computing*, 2(2):153–163, 1981. MR0622712

Sara Anna van de Geer. *Adaptive quantile regression*. University of Leiden. Mathematical Institute, 2003. MR2498244

Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Rebecca M Willett and Robert D Nowak. Multiscale poisson intensity and density estimation. *IEEE Transactions on Information Theory*, 53(9):3171–3187, 2007. MR2417680

Sascha Wirges, Tom Fischer, Christoph Stiller, and Jesus Balado Frias. Object detection and classification in occupancy grid maps using deep convolutional

networks. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3530–3535. IEEE, 2018.

Steven Siwei Ye and Oscar Hernan Madrid Padilla. Non-parametric quantile regression via the k-nn fused lasso. *Journal of Machine Learning Research*, 22(111):1–38, 2021. MR4279762

Keming Yu and Rana A Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001. MR1861390