Adaptive Estimation via Optimal Decision Trees

Subhajit Goswami

Tata Institute of Fundamental Research (TIFR), Mumbai

ICTS Program on Advances in Applied Probability

Joint work with Sabyasachi Chatterjee, UIUC

January 6, 2021

Dyadic CART

The Annals of Statistics
1997, Vol. 25, No. 5, 1870–1911

CART AND BEST-ORTHO-BASIS: A CONNECTION[1]

BY DAVID L. DONOHO

Stanford University and University of California, Berkeley

We study what we call "dyadic CART"—a method of nonparametric regression which constructs a recursive partition by optimizing a complexity penalized sum of squares, where the optimization is over all recursive partitions arising from midpoint splits. We show that the method is adaptive to unknown degrees of anisotropic smoothness. Specifically, consider the anisotropic smoothness classes of Nikol'skii, consisting of bivariate functions $f(x_1, x_2)$ whose finite difference of distance h in direction i is bounded in L^p norm by Ch^{δ_i}, $i = 1, 2$. We show that dyadic CART, with an appropriate complexity penalty parameter $\lambda \sim \sigma^2 \cdot \text{Const} \cdot \log(n)$, is within logarithmic terms of minimax over every anisotropic smoothness class $0 < C < \infty$, $0 < \delta_1, \delta_2 \leq 1$.

The proof shows that dyadic CART is identical to a certain adaptive best-ortho-basis algorithm based on the library of all anisotropic Haar bases. Then it applies empirical basis selection ideas of Donoho and Johnstone. The basis empirically selected by dyadic CART is shown to be

- Dyadic CART (DC) was introduced in Donoho (1997) as a nonparametric regression method for denoising $2D$ signals.

- It computes a globally optimal dyadic decision tree based on solving a penalized least squares optimization problem.

- It was shown that DC is minimax rate optimal over bivariate function classes which show anisotropic smoothness.

- Ideas related to Dyadic CART can be useful in estimating spatially heterogenous nonparametric function classes of recent interest.

    1. Rectangular Piecewise Constant or Polynomial Functions in General Dimensions.
    2. 2D Bounded Variation Images/Matrices (alternative to TV Denoising).
    3. Univariate Bounded Variation Functions of General Orders (alternative to Trend Filtering).

Optimally Adaptive Estimation of Piecewise Constant Signals

Setup

There is an underlying signal $\theta^* \in \mathbb{R}^{L_{d,n}}$ where $L_{d,n} := \{1, \ldots, n\}^d$.

We observe

$$y = \theta^* + \sigma\epsilon$$

where $\sigma > 0$ is noise strength and ϵ is an array consisting of mean 0 independent subgaussian random variables.

We measure the performance of an estimator $\hat{\theta}$ in terms of the mean squared error:

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

where $N := n^d$ is the sample size and $\|\cdot\|$ is Euclidean norm.

Definition of piecewise constant

A rectangle $R \subset L_{d,n}$ is simply a product of disjoint discrete intervals $\otimes_{i=1}^{d}[a_i, b_i]$.

A rectangular partition of $R \subset L_{d,n}$ is a set of disjoint rectangles $\Pi = \{R_1, \ldots, R_k\}$ whose union is the whole of R.

An array $\theta^* \in \mathbb{R}^{L_{d,n}}$ is said to be piecewise constant if there exists a rectangular partition $\Pi = \{R_1, \ldots, R_k\}$ of $L_{d,n}$ such that θ_{R_i} is constant.

Let us define $k_{\text{all}}(\theta^*)$ to be the cardinality of the minimal rectangular partition Π such that θ^* is constant on the *blocks* of Π.

We say θ^* is piecewise constant with k pieces if $k_{\text{all}}(\theta^*) = k$.

Minimax risk

An oracle, who knows the rectangular partition on the blocks of which θ^* is constant, will just fit the mean of the observation array y within each of the blocks.

This means the oracle estimator incurs $\sigma^2 \frac{k_{\text{all}}(\theta^*)}{N}$ as MSE

A standard argument gives the following lower bound:

$$\inf_{\hat{\theta}} \sup_{\theta^* \in \mathbb{R}^{L_d, n} : k_{\text{all}}(\theta^*) \leq k} \text{MSE}(\hat{\theta}, \theta^*) \geq C_d \sigma^2 \frac{k}{N} \log \frac{N}{k}.$$

where C_d is a constant that can depend on d.

Main question

Q: Does there exist an estimator which, in all dimensions,

- *attains the minimax rate risk bound* $O(\sigma^2 \frac{k_{\text{all}}(\theta^*)}{N} \log N)$ *adaptively for all* θ^*.
- *is possible to compute in polynomial time in the sample size* N*?*

Recently, estimators based on convex optimization such as TV Denoising in Chatterjee and Goswami (2020) and Hardy Krause Variation Denoising Ortelli and Van De Geer(2019), Fang, Guntuboyina and Sen(2020) have been studied in the above context.

To the best of our knowledge, the above question has not been rigorously answered in the literature. We will propose a solution in this talk.

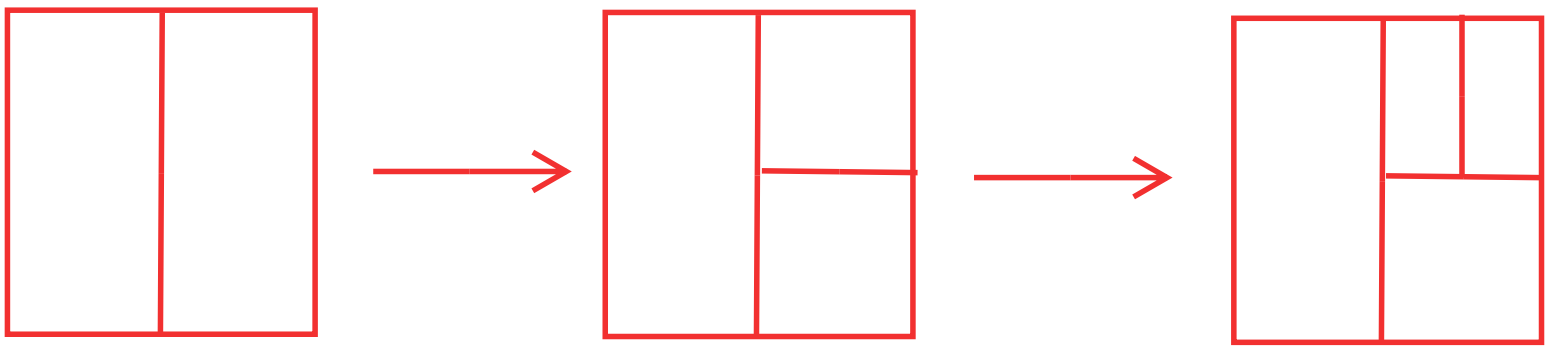
Dyadic CART: definition and main result

Recursive dyadic partitions

A dyadic split of a rectangle R is to choose any of its d sides and divide it in half. This split produces two rectangles of (almost) equal size.

Starting from the trivial partition $L_{d,n}$ we can recursively generate finer partitions by performing dyadic splits.

A recursive dyadic partition (RDP) is any partition reachable by such successive binary divisions.



Dyadic CART

Let us denote the set of all recursive dyadic partitions of $L_{d,n}$ as $\mathcal{P}_{\mathrm{rdp},d,n} =: \mathcal{P}_{\mathrm{rdp}}$.

We can now consider the following minimal partitioning optimization problem:

$$\hat{\Pi}_{\mathrm{rdp},\lambda} := \underset{\Pi \in \mathcal{P}_{\mathrm{rdp}}}{\operatorname{argmin}} \left(\|y - \Pi y\|^2 + \lambda |\Pi| \right).$$

The estimator

$$\hat{\theta}_{\mathrm{rdp},\lambda} := \hat{\Pi}_{\mathrm{rdp},\lambda} y$$

takes the form of a piecewise constant array on the partition $\hat{\Pi}_{\mathrm{rdp}}$. This estimator is precisely the Dyadic CART estimator which was proposed in Donoho (1997).

The optimization problem can be solved in at most $O_d(N)$ basic operations by dynamic programming.

Main result for Dyadic CART

For any array $\theta \in R^{L_{d,n}}$ define $k_{\mathrm{rdp}}(\theta)$ to be the cardinality of the minimal rectangular recursive dyadic partition Π such that θ is constant on the blocks of Π. By definition, the following inequality holds for any array θ;

$$k_{\mathrm{all}}(\theta) \leq k_{\mathrm{rdp}}(\theta).$$

Theorem (Oracle risk bound for Dyadic CART)

There exists a constant C such that if we set $\lambda \geq C\sigma^2 \log N$ then for any $\theta^* \in \mathbb{R}^N$ we have the following risk bound:

$$\mathrm{MSE}(\hat{\theta}_{\mathrm{rdp},\lambda}, \theta^*) \leq \inf_{\theta \in \mathbb{R}^{L_{n,d}}} \left(2\,\frac{\|\theta - \theta^*\|^2}{N} + \frac{\lambda k_{\mathrm{rdp}}(\theta)}{N}\right).$$

How big is $k_{\mathrm{rdp}}(\theta)$ compared to $k_{\mathrm{all}}(\theta)$?

When $d = 1$ or 2, any rectangular partition of $L_{d,n}$ can be refined into a RDP increasing the number of rectangles by a poly-logarithmic factor.

$$k_{\mathrm{rdp}}(\theta) \leq k_{\mathrm{all}}(\theta) \;\; C(\log n)^d \;\; \forall \;\; \theta \in \mathbb{R}^{L_{d,n}}$$

In particular, we can conclude under our choice of λ and $d = 1$ or 2,

$$\mathrm{MSE}(\hat{\theta}_{\mathrm{rdp},\lambda}, \theta^*) \leq C_d \sigma^2 (\log N)^{d+1} \frac{k_{\mathrm{all}}(\theta)}{N}.$$

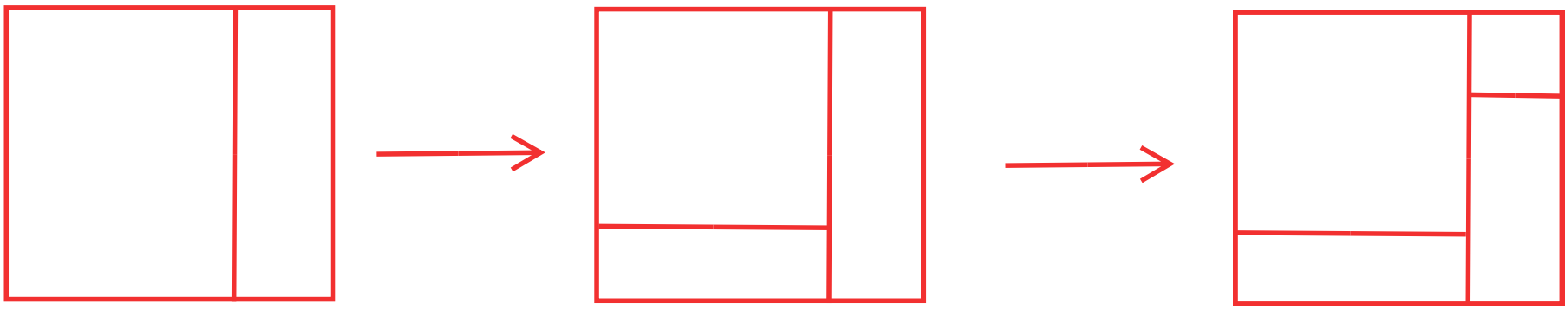
Optimal Regression Tree: definition and main result

Hierarchical partitions / decision trees

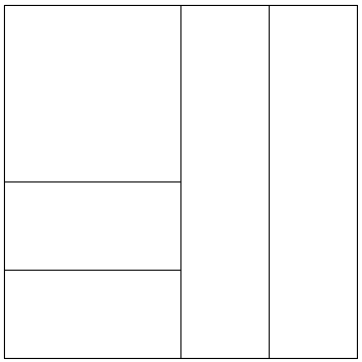
A hierarchical split of a rectangle R is to choose any of its d sides and divide it along some point. This split produces two rectangles of not necessarily equal size.

Starting from $L_{d,n}$ we can now recursively generate finer partitions by performing hierarchical splits.

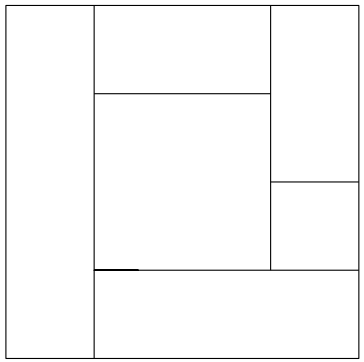
A hierarchical partition/decision tree is any partition reachable by such successive hierarchical splits.



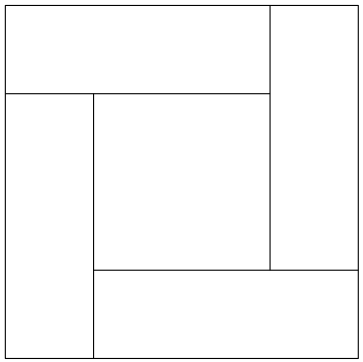
Not all partitions are Hierarchical



(a)



(b)



(c)

Image (a) is an example of a recursive dyadic partition of the plane. Image (b) is non-dyadic but is a hierarchical partition. Image (c) is an example of a non hierarchical partition. An easy way to see this is that there is no split from top to bottom or left to right.

Optimal Regression Tree

Let us denote the set of all hierarchical partitions/decision trees of $L_{d,n}$ as $\mathcal{P}_{\text{tree},d,n} =: \mathcal{P}_{\text{tree}}$.

We can now consider the corresponding estimator which optimizes over all partitions in $\mathcal{P}_{\text{tree}}$.

$$\hat{\Pi}_{\text{tree},\lambda} := \underset{\Pi \in \mathcal{P}_{\text{tree}}}{\operatorname{argmin}} \left(\|y - \Pi y\|^2 + \lambda |\Pi| \right).$$

The estimator

$$\hat{\theta}_{\text{tree},\lambda} := \hat{\Pi}_{\text{tree},\lambda} y$$

takes the form of a piecewise constant array on the partition $\hat{\Pi}_{\text{tree},\lambda}$.

Crucially, ORT can be computed by $O(n^{(2d+1)} = N^{(2+1/d)})$ basic operations.

Main result for Optimal Regression Tree

For any array $\theta \in R^{L_{d,n}}$ define $k_{\text{tree}}(\theta)$ to be the cardinality of the minimal hierarchical partition / decision tree Π such that θ is constant on the blocks of Π. By definition, the following inequality holds for any array θ;

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

Theorem

There exists a universal constant C such that if we set $\lambda \geq C\sigma^2 \log N$, then for any $\theta^* \in \mathbb{R}^N$ we have the following risk bound:

$$\text{MSE}(\hat{\theta}_{\text{tree},\lambda}, \theta^*) \leq \inf_{\theta \in \mathbb{R}^{L_{n,d}}} \left(2\,\frac{\|\theta - \theta^*\|^2}{N} + \frac{\lambda k_{\text{tree}}(\theta)}{N}\right).$$

Comparing $k_{\text{tree}}(\theta)$ to $k_{\text{all}}(\theta)$: some results from Computational Geometry

In 2 dimensions, any rectangular partition can be refined into a decision tree where the number of rectangular pieces at most doubles.

Theorem (Berman et al (2002))

Given any partition $\Pi \in \mathcal{P}_{\text{all},2,n}$ *there exists a refinement* $\tilde{\Pi} \in \mathcal{P}_{\text{tree},2,n}$ *such that* $|\tilde{\Pi}| \leq 2|\Pi|$. *As a consequence, for any matrix* $\theta \in \mathbb{R}^{n \times n}$, *we have*

$$k_{\text{tree}}(\theta) \leq 2k_{\text{all}}(\theta).$$

In particular, when $d = 2$ (and of course $d = 1$), we can conclude under our choice of λ,

$$\text{MSE}(\hat{\theta}_{\text{tree},\lambda}, \theta^*) \leq C\sigma^2(\log N)\frac{k_{\text{all}}(\theta^*)}{N}.$$

Comparing $k_{\text{tree}}(\theta)$ to $k_{\text{all}}(\theta)$: some results from Computational Geometry

The aspect ratio of a rectangle is the ratio of its maximum sidelength to its minimum sidelength.

A rectangular partition $\Pi \in \mathcal{P}_{\text{all},d,n}$ is called α-fat ($\alpha \geq 1$) if all its blocks have aspect ratio bounded by α.

Theorem (De Berg (1995))

Given any partition $\Pi \in \mathcal{P}_{\text{all},d,n}$ which is α fat, there exists a refinement $\tilde{\Pi} \in \mathcal{P}_{\text{tree},d,n}$ such that $|\tilde{\Pi}| \leq C_{d,\alpha}|\Pi|$.

In particular, when all the rectangular level-sets of θ^* are α-fat, we have under our choice of λ,

$$\text{MSE}(\hat{\theta}_{\text{tree},\lambda}, \theta^*) \leq C_{d,\alpha}\sigma^2(\log N)\frac{k_{\text{all}}(\theta^*)}{N}.$$

Summary of main results

- DC attains the risk bound $O(\frac{k_{all}(\theta^*)}{N}(\log N)^{d+1})$ for all θ^* in dimensions $d = 1, 2$.
- ORT attains the improved risk bound $O(\frac{k_{all}(\theta^*)}{N}(\log N))$ for all θ^* in dimensions 2 (and 1).
- ORT attains the improved risk bound $O(\frac{k_{all}(\theta^*)}{N}(\log N))$ for all θ^* which are piecewise constant on a fat partition in all dimensions.
- DC and ORT computable in $O_d(N), O_d(N^{2+1/d})$ basic operations.

Computation

For any rectangle $R \subset L_{d,n}$ we can define the corresponding subproblem restricted to R.

$$\text{Problem}(R) := \min_{\Pi:\Pi\in\mathcal{P}_{hier},R} (\|y_R - \Pi y_R\|^2 + K|\Pi|).$$

The key point is that the optimization satisfies a recurrence relation:

$$\text{Problem}(R) = \min_{R_1,R_2:R_1\cup R_2=R} (\text{Problem}(R_1) + \text{Problem}(R_2), \|y_R - \bar{y}_R\|^2 + K).$$

$$\begin{aligned}\text{Complexity} &= |\textit{Rectangles}| \times |\textit{Splits for each rectangle}| \\ &= O(N^2) \times O(n) = O(N^2\ n).\end{aligned}$$

Estimation of Bounded Variation Functions in 2D

2D Total Variation Denoising

Think of $L_{2,n}$ as the 2 dimensional regular $n \times n$ lattice graph. Define

$$\mathrm{TV}(\theta) := \frac{1}{n} \sum_{(u,v) \in E_{2,n}} |\theta_u - \theta_v|$$

where $E_{2,n}$ is the edge set of the graph $L_{2,n}$.

The 2D TVD estimator was first introduced by Rudin(1992) for image denoising.

$$\hat{\theta}_{\mathrm{TVD},\lambda} := \underset{\theta \in \mathbb{R}^{L_{2,n}}}{\operatorname{argmin}} \|y - \theta\|^2 + \lambda \mathrm{TV}(\theta).$$

Minimax Rates

Hutter Rigollet(2016), Sadhanala et al(2016) show that a well tuned TVD estimator is nearly (up to log factors) minimax rate optimal over the class $\{\theta \in \mathbb{R}^{L_{2,n}} : \mathrm{TV}(\theta) \leq V\}$ attaining MSE bounded by $\tilde{O}(\frac{V}{\sqrt{N}})$.

Theorem (Risk Bound for Dyadic CART)

Let $\theta^* \in \mathbb{R}^{L_{2,n}}$ be the underlying truth and let $V^* := TV(\theta^*)$. The following risk bound holds for the Dyadic CART estimator $\hat{\theta}_{\mathrm{rdp},\lambda}$ with $\lambda \geq C\sigma^2 \log N$ where C is an absolute constant.

$$\mathrm{MSE}(\hat{\theta}_{\mathrm{rdp},\lambda}, \theta^*) \leq C\sigma \frac{V^*}{\sqrt{N}} \log N.$$

Main result for Dyadic CART

For any array $\theta \in R^{L_{d,n}}$ define $k_{\mathrm{rdp}}(\theta)$ to be the cardinality of the minimal rectangular recursive dyadic partition Π such that θ is constant on the blocks of Π. By definition, the following inequality holds for any array θ;

$$k_{\mathrm{all}}(\theta) \leq k_{\mathrm{rdp}}(\theta).$$

Theorem (Oracle risk bound for Dyadic CART)

There exists a constant C such that if we set $\lambda \geq C\sigma^2 \log N$ then for any $\theta^* \in \mathbb{R}^N$ we have the following risk bound:

$$\mathrm{MSE}(\hat{\theta}_{\mathrm{rdp},\lambda}, \theta^*) \leq \inf_{\theta \in \mathbb{R}^{L_{n,d}}} \left(2\,\frac{\|\theta - \theta^*\|^2}{N} + \frac{\lambda k_{\mathrm{rdp}}(\theta)}{N}\right).$$

Approximation Result

Proposition

Let $\theta \in \mathbb{R}^{L_{2,n}}$ *be such that* $nTV(\theta) = V$. *For any* $\delta > 0$, *there exists a* $\theta^{'} \in \mathbb{R}^{n}$ *such that*

a) $k_{\mathrm{rdp}}(\theta^{'}) \simeq V\delta^{-1}$ *and*

b) $\|\theta - \theta^{'}\|^2 \simeq V\delta$

Sketch of proof of the main results

Sketch of proof: formulation in terms of Union of Subspaces

First notice that the parameter space Θ in both cases can be written as $\cup_{S \in \mathcal{S}} S$ where $\mathcal{S}$ is a finite collection of subspaces of $\mathbb{R}^N$.

Therefore our estimators are special cases of the penalized least squares estimator

$$\hat{\theta}_\lambda := \underset{\theta \in \Theta}{\operatorname{argmin}} \left(\|y - \theta\|^2 + \lambda k_{\mathcal{S}}(\theta) \right),$$

where $\lambda \geq 0$ is the tuning parameter and

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

Secondly, notice that $N_k(\mathcal{S}) := |S : \operatorname{Dim}(S) = k, S \in \mathcal{S}| \leq N^k$ for every $k \in [N]$.

Sketch of proof: a generic result

Under the setting described in the previous slide, one can prove:

Theorem (Union of Subspaces)

Setting $\lambda \geq C\sigma^2 \log N$ *and a constant* $C > 0$ *depending solely on* c*, we have*

$$\mathbb{E}\|\hat{\theta}_\lambda - \theta^*\|^2 \leq \inf_{\theta \in \Theta} \left[2\|\theta - \theta^*\|^2 + \lambda k_{\mathcal{S}}(\theta)\right] + C\sigma^2.$$

The MSE bounds for both of our estimators now follow directly from this result.

Outline for the proof of the generic result

We assume $y = \theta^* + \sigma Z$ where Z has i.i.d. standard Gaussian entries

Now use

$$\|y - \hat{\theta}\|^2 + \lambda\, k_{\mathcal{S}}(\hat{\theta}) \leq \|y - \theta\|^2 + \lambda\, k_{\mathcal{S}}(\theta)\ \forall\, \theta \in \mathcal{S}\,(\text{by definition})$$

and expand the squares after substituting $\theta^* + \sigma Z$ for y to obtain

$$\|\hat{\theta} - \theta^*\|^2 \leq \|\theta - \theta^*\|^2 + \lambda k_{\mathcal{S}}(\theta) + (2\sigma\, \langle Z, \hat{\theta} - \theta\rangle - \lambda k_{\mathcal{S}}(\hat{\theta})).$$

After some routine manipulation we get

$$\|\theta^* - \hat{\theta}\|^2 \leq 2\|\theta^* - \theta\|^2 + 2\lambda k_{\mathcal{S}}(\theta) + L(Z)$$

where, with $\overline{v}$ denoting $\frac{v}{\|v\|}$,

$$L(Z) := C_1\sigma^2\langle\, Z, \overline{\hat{\theta} - \theta}\rangle^2 - C_2\lambda k_{\mathcal{S}}(\hat{\theta}).$$

Outline for the proof of the generic result

With $\overline{v}$ denoting $\frac{v}{\|v\|}$,

$$L(Z) := C_1\sigma^2\langle Z, \overline{\hat{\theta}-\theta}\rangle^2 - C_2\lambda k_{\mathcal{S}}(\hat{\theta}).$$

Therefore, we can write

$$L(Z) \leq \max_{k\in[n]}\Big[C_1\sigma^2 \sup_{S\in\mathcal{S}:\mathrm{Dim}(S)=k}\ \sup_{v\in S}\langle Z, \overline{v-\theta}\rangle^2 \ - \ C_2\lambda k\Big].$$

However, using fairly standard arguments one can show

$$\mathbb{E}\sup_{v\in S,\, v\neq\theta}\langle Z, \overline{v-\theta}\rangle \leq (\mathrm{Dim(S)}^{1/2}+1).$$

Outline for the proof of the generic result

We can write

$$L(Z) \leq \max_{k \in [n]} \Big[C_1 \sigma^2 \sup_{S \in \mathcal{S}: \mathrm{Dim}(S) = k} \sup_{v \in S} \langle Z, \overline{v - \theta} \rangle^2 \; - \; C_2 \lambda k \Big].$$

Using fairly standard arguments one can show

$$\mathbb{E} \sup_{v \in S,\, v \neq \theta} \langle Z, \overline{v - \theta} \rangle \leq (\mathrm{Dim}(\mathrm{S})^{1/2} + 1).$$

Hence the Gaussian concentration inequality allows us to deduce for any $S \in \mathcal{S}$ with $\mathrm{Dim}(S) = k$,

$$\mathbb{P}\Big(\Big[\sup_{v \in S} \langle Z, \overline{v - \theta} \rangle^2 - \frac{C_2 \lambda}{C_1 \sigma^2} k \Big] > t \Big) \leq C \exp\Big(- c' \Big[t + \frac{\lambda}{\sigma^2} k \Big] \Big) \;\; \forall t \geq 0.$$

Outline for the proof of the generic result

We can write

$$L(Z) \leq \max_{k\in[n]} \Big[C_1\sigma^2 \sup_{S\in\mathcal{S}:\mathrm{Dim}(S)=k} \sup_{v\in S} \langle Z, \overline{v-\theta}\rangle^2 \; - \; C_2\lambda k \Big].$$

We can deduce for any $S \in \mathcal{S}$ with $\mathrm{Dim}(S) = k$,

$$\mathbb{P}\Big(\Big[\sup_{v\in S} \langle Z, \overline{v-\theta}\rangle^2 - \frac{C_2\lambda}{C_1\sigma^2} k \Big] > t \Big) \leq C \exp\Big(- c' \Big[t + \frac{\lambda}{\sigma^2} k \Big] \Big) \;\; \forall t \geq 0.$$

Now using a (double) union bound we get for all $t \geq 0$,

$$\mathbb{P}(L(Z) > t) \leq C \sum_{k\in[n]} N_k(\mathcal{S}) \exp\Big(- \frac{c'}{\sigma^2} \big[t + \lambda k \big] \Big).$$

Outline for the proof of the generic result

We can write

$$L(Z) \leq \max_{k \in [n]} \Big[C_1 \sigma^2 \sup_{S \in \mathcal{S}: \mathrm{Dim}(S) = k} \sup_{v \in S} \langle Z, \overline{v - \theta} \rangle^2 \; - \; C_2 \lambda k \Big].$$

We have for all $t \geq 0$,

$$\mathbb{P}(L(Z) > t) \leq C \sum_{k \in [n]} N_k(\mathcal{S}) \exp\Big(- \frac{c'}{\sigma^2} \big[t + \lambda k \big] \Big).$$

The result now follows by integrating this bound with the choice $\lambda \geq C \, \sigma^2 \, \log N$ upon recalling that

$$\log N_k(\mathcal{S}) \leq ck \log N.$$

Concluding Remarks

- One can define versions of DC and ORT which fits piecewise polynomials on rectangles.
- The oracle inequalities for DC and ORT can be used to show minimax rate optimality for other function classes (possibly yet to be explored) using appropriate approximation theoretic results.
- DC/ORT has a natural computable extension to random design.
- Can λ can be chosen in a data driven way so that all the present guarantees are maintained?

S. Chatterjee and S. Goswami. **Adaptive Estimation of Multivariate Piecewise Polynomials and Bounded Variation Functions by Optimal Decision Trees.** To appear in *Annals of Statistics*. Available at https://arxiv.org/pdf/1911.11562.pdf

THANK YOU FOR YOUR ATTENTION!