Journal of Machine Learning Research 23 (2022) 1-XX Submitted 1/21; Revised 5/22; Published 9/22

Interpretability May Require Rethinking Accuracy: An Interplay of Depth and Optimization

Akash Kumar AKK002@UCSD.EDU
Department of Computer Science & Engineering
University of California-San Diego
La Jolla, CA 92093-0404, USA

Editor: My editor

Abstract

In this work, we study the interplay between interpretability and accuracy in machine learning models by comparing decision trees and neural networks. While decision trees are structurally interpretable due to their axis-aligned, rule-based partitions, shallow ReLU networks offer expressive power but lack transparency.

From the point of view of interpretability, we show that there is an inherent dichotomy in approximation capabilities of neural networks depending on if a problem is viewed from the lens of regression or classification. For approximating decision trees as solving a classification problem is rather easy but it comes at the cost of interpretability. If we consider the average of classifier's jacobian (gradient with respect to the input) close to the decision boundary then a model optimal for classification could still exhibit high rank albeit requires a smaller one for interpretable features corresponding to the simple decision boundary. We study approximation of decision trees on three axes: loss function, depth and width of a network. First, we show hard-threshold decision trees lie outside the RBV class by proving their Radon Total Variation (RTV) norm is infinite. We then analyze three natural smoothing schemes—continuous piecewise linear (CPwL), sigmoidal, and Gaussian—and find that both CPwL and sigmoid smoothings still yield infinite Radon Total Variation, whereas Gaussian smoothing results in finite RTV with explicit dependence on the input dimension d.

With depth, you can approximate these simple boundaries in theory arbitrarily well but in order to learn them we need the right choice of optimizers and loss functions – 1) sgd learn these boundaries better than adam, 2) width is detrimental and 3) harmonic and cross entropy loss could be beneficial over square loss.

Keywords: Interpretability, Shallow Networks, Decision Trees, Function Approximation, Neural Networks List of keywords

1 Introduction

In safety–critical and socially–sensitive settings, we need models whose behavior we can *explain* and audit. Yet the highest accuracies are still delivered by opaque neural networks, fueling a debate about whether an interpretability–accuracy trade-off is fundamental or merely an artifact of modelling choices (Rudin, 2019; Doshi-Velez and Kim, 2017; Atrey et al., 2025).

A canonical interpretable baseline is the *axis-aligned decision tree*. It classifies $x \in \mathbb{R}^d$ by a sequence of one-dimensional threshold tests; algebraically, it induces a piecewise-constant function $f_{\mathsf{DT}} = \mathbf{1}_A$ where $A \subset \mathbb{R}^d$ is a finite union of axis-aligned boxes.

In contrast, shallow neural networks (e.g., single-hidden-layer) combine multiple ridge functions:

$$\boldsymbol{x} \mapsto \sum_{k=1}^{K} v_k \, \sigma(\boldsymbol{w}_k^\top \boldsymbol{x} - b_k),$$

where K is the network width, $v_k \in \mathbb{R}$, $\boldsymbol{w}_k \in \mathbb{R}^d \setminus \{0\}$ are weights, and $b_k \in \mathbb{R}$ are biases. These networks are naturally analyzed in the infinite-width (ridgelet) limit via the *Radon total variation* (RTV) seminorm $\|f\|_{\mathcal{R}}$ (Savarese et al., 2019; Ongie et al., 2020). Intuitively, $\|f\|_{\mathcal{R}}$ upper-bounds the amount and sharpness of level-set bending and serves as a geometric proxy for model complexity (and thus a candidate proxy for interpretability). See §3 for the precise definition and its equivalence to a minimum-mass ridgelet representation.

Two tasks that are often conflated. There are (at least) two ways to "learn a tree" with a shallow net: (i) *classification only*: learn a score $s : \mathbb{R}^d \to [0, 1]$ whose thresholding yields the correct set A; and (ii) *score learning / regression*: learn a score that is *calibrated* and *close* to $\mathbf{1}_A$ (e.g., in $\mathrm{L}^1(P)$), so that its level sets and gradients align with the symbolic features (splits) of the tree. The former is easier and common in practice (thresholding the logits), but it can hide the learned features: many very different scores lead to the *same* thresholded classifier.

This paper. We formalize the distinction above and show that, for shallow networks, the *classification problem*—learning a score whose thresholding yields the correct decision set—is easy, since thresholding a suitably constructed score function can exactly recover the decision set without requiring complex function approximation. In contrast, the *regression problem*—learning a score that is calibrated and close to the symbolic tree function—faces a principled complexity–accuracy frontier. Our lens is the RTV seminorm. We prove that the hard tree $\mathbf{1}_A$ has infinite RTV (§3), that several naive smoothings retain infinite RTV, and that a carefully chosen smoothing admits *finite* RTV with an explicit rate. Crucially, this smoothing has an *exact threshold* that recovers A, yet its *closeness* to $\mathbf{1}_A$ necessarily degrades as RTV is forced smaller. Experiments confirm that different optimizers populate different portions of the empirical frontier.

Contributions (informal). Let A be a finite union of axis-aligned boxes in $\mathbb{R}^d$.

1. (*Hard tree is not representable at bounded RTV.*) $\|\mathbf{1}_A\|_{\mathcal{R}} = +\infty$ for all $d \geq 1$; exact trees sit outside any bounded-RTV ball.

2. (*Classification is easy.*) We exhibit a sharp barrier score s_B whose threshold at 1 *exactly* recovers A and has *finite* RTV. Under a mild tube-mass assumption on the data distribution P near ∂A, we further obtain $E[|s_k(X) - \mathbf{1}_A(X)|] \lesssim (dc_0)^\beta \, \lambda^{-\beta} \; + \; e^{dc_0} \, C \, \Gamma(\beta + 1) \, \lambda^{-\beta}$.

3. (*Experiments.*)[**A:** provide all experiments] On synthetic unions of rectangles, we trace the frontier between $\mathrm{L}^1(P)$ gap to $\mathbf{1}_A$ and a practical RTV proxy, and show optimizer-dependent biases in where training lands on this frontier (with or without weight decay).

So, if one only cares about the *thresholded classifier*, shallow nets can represent trees easily. If one cares about learning a score that is both *interpretable* (low RTV, gradients aligned to splits) and *close* to the symbolic model, a quantitative trade-off emerges. This clarifies what exactly is in tension and why thresholding alone can mask feature misalignment.

2 Related Work

The accuracy–interpretability debate Early empirical studies reported an apparent performance gap between transparent models (linear regressions, GAMs, decision trees) and deep neural nets (Doshi-Velez and Kim, 2017). More recent domain-specific benchmarks nuance this view—for instance, carefully tuned interpretable models can rival black-box baselines on extreme-event prediction (Lovo et al., 2025), while recent case studies report a nuanced, non-monotonic relationship between interpretability and predictive performance: as interpretability decreases, accuracy often improves, but there are notable instances where more interpretable models perform comparably to or even outperform less interpretable ones (Atrey et al., 2025). Critics even label the trade-off a "myth" in high-stakes settings (Rudin, 2019). Our analysis sharpens the discussion by identifying a dimension-dependent regime—captured through a geometric complexity measure—where high accuracy necessarily coincides with large complexity, thereby clarifying why empirical findings sometimes diverge.

Complexity measures for neural functions Generalisation guarantees for neural networks are typically phrased in weight-space norms—e.g., the path-norm (Neyshabur et al., 2015), products of spectral norms (Bartlett et al., 2017), or Neural Tangent Kernel radii (Jacot et al., 2020). While these quantities correlate with test error—and, in some cases, adversarial robustness—their connection to human interpretability remains indirect. A complementary viewpoint is provided by the *Radon bounded-variation* space $\mathcal{R}\mathrm{BV}^2(\Omega)$ ($\Omega \subseteq \mathbb{R}^d$), defined by bounded Radon-domain total variation (RTV) (Savarese et al., 2019). Representer theorems show that shallow ReLU networks trained with weight decay lie in $\mathcal{R}\mathrm{BV}^2(\Omega)$ (Parhi and Nowak, 2021), and recent embedding results connect Sobolev spaces to RBV, clarifying when sharp or low-dimensional structures are captured (Ongie et al., 2020; Mao et al., 2024). Building on this framework, we extend RTV analysis to both smoothed and hard decision-tree limits, illuminating regimes in which symbolic transparency fundamentally clashes with predictive accuracy.

Approximation behaviour of shallow *versus* deep networks It is well-known that depth-2 neural networks are universal approximators of any continuous function on a bounded domain when equipped with reasonable activation functions (Cybenko, 1989; Hornik et al., 1989; Funahashi, 1989). In the case of shallow networks with a single hidden layer, it is established that approximating a C^n-function on a d-dimensional set with infinitesimal error ϵ requires a network of size roughly $\epsilon^{-d/n}$, assuming a smooth activation function (see, e.g., (Pinkus, 1999; Mhaskar, 1996) for further details). This result is refined in Yarotsky (2018), which analyzes both shallow and deep ReLU networks, demonstrating how depth circumvents the curse of dimensionality inherent to shallow architectures.

Discontinuous targets pose far greater challenges: even the *smoothed* indicators produced by decision trees exhibit a Radon total-variation ($\mathcal{R}\mathrm{TV}^2$) that grows exponentially with the ambient dimension. Consequently, any predictor constrained to bounded RTV would still require exponentially many parameters—or additional depth—to match tree-level accuracy.

Depth fundamentally alters this narrative. Telgarsky (2016) constructs networks with $\Theta(k^3)$ layers and $\mathcal{O}(1)$ width that *cannot* be approximated by $\mathcal{O}(k)$-layer networks unless their size swells to $\Omega(2^k)$ units; this separation extends to boosted decision trees, which would need $\Omega(2^{k^3})$ nodes to approximate the same target. Related depth–width separations appear in Eldan and Shamir (2016)

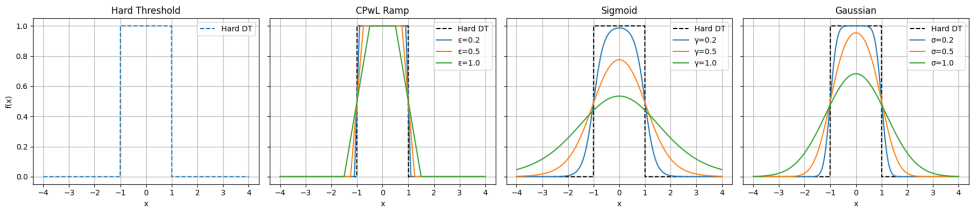


Figure 1: Illustration of the three smoothing schemes (ramp, sigmoidal, Gaussian) applied to a simple axis-aligned threshold.

and subsequent work, where certain three-layer ReLU networks are shown to outperform all two-layer alternatives of polynomial width.

Taken together with our RTV calculations, these results paint a coherent picture: sharp, axis-aligned decision boundaries force shallow, bounded-complexity networks to explode in size, whereas deeper architectures can trade width for depth to maintain polynomial scaling in both d and ϵ^{-1}.

Smooth decision trees and differentiable forests Gradient–based tree learning has grown into a sizeable literature beyond Kontschieder et al. (2015); Frosst and Hinton (2017). Neural Oblivious Decision Ensembles (NODE) couple stacked, feature-shared "oblivious" trees with end-to-end back-propagation and set a strong tabular benchmark (Popov et al., 2019). Adaptive Neural Trees dynamically *grow* the topology while learning leaf predictors (Tanno et al., 2019). The *Tree Ensemble Layer* slots an entire soft forest into a deep network and trains it jointly with the upstream representation (Hazimeh et al., 2020). More recently, GRANDE optimises large differentiable forests with Adam and outperforms gradient-boosting on many datasets (Marton et al., 2024). A parallel thread explores fully differentiable *oblique* splits, e.g. training axis-free trees by vanilla gradient descent (Panda et al., 2024). Collectively, these works confirm that back-prop can fit high-capacity trees, yet our RTV calculation shows that merely widening the routing band leaves the intrinsic complexity exponential in dimension, clarifying why interpretability gains remain elusive at scale.

3 Problem Setup

Data space. We work on the ambient Euclidean space $\mathcal{X} = \mathbb{R}^d$ with $d \geq 1$. We denote a datapoint by $\boldsymbol{x} \in \mathcal{X}$ and scalar values by $y, z, \omega \in \mathbb{R}$.

Axis-aligned decision trees. A depth–D axis-aligned tree is determined by ordered splits $(\mathbf{w}_i, b_i)_{i=1}^{D}$ where $\mathbf{w}_i \in \{\pm \mathbf{e}_{j_i}\}$ selects the coordinate $j_i \in [d]$ and $b_i \in \mathbb{R}$ is the threshold. The classifier is the indicator

$$f_{\mathsf{DT}}(\boldsymbol{x}) \;=\; 1\{\boldsymbol{x} \in A\}, \qquad A = \bigcup\nolimits_k B_k$$

with A a union of axis-aligned boxes $\{B_k\}$ (corresponding to the leaves). Its decision boundary is a union of coordinate-aligned $(d-1)$-planes, yielding perfect interpretability and $O(D)$ evaluation cost. However, the jump discontinuities place f_{DT} outside smooth function classes such as reproducing kernel Hilbert spaces (RKHS) or the space of 2-layered infinite width ReLU neural networks, giving $\|f_{\mathsf{DT}}\|_{\mathcal{R}} = \infty$ (see Section 4). For analytic control we therefore introduce smooth surrogates that retain the tree structure but soften each split (see Fig. 1 for illustration).

Smoothed decision trees. All surrogates keep the same $(\mathbf{w}_i, b_i)$ and depth D; they differ only in how the sign test $1\{\mathbf{w}_i^\top \boldsymbol{x} + b_i > 0\}$ is replaced.

Piecewise-linear ramp smoothing. For a margin width $\epsilon > 0$ define

$$\rho_\epsilon(z) = \begin{cases} 0, & z \le -\frac{\epsilon}{2}, \\ \frac{z}{\epsilon} + \frac{1}{2}, & |z| < \frac{\epsilon}{2}, \\ 1, & z \ge \frac{\epsilon}{2}. \end{cases} \tag{1}$$

The ramp-smoothed tree is

$$f_{\mathsf{DT},\epsilon}(\boldsymbol{x}) \;=\; \prod_{i=1}^{D} \rho_\epsilon\big(\mathbf{w}_i^\top \boldsymbol{x} + b_i\big). \tag{2}$$

It coincides with f_{DT} outside width-ϵ slabs and converges to the hard tree as $\epsilon \to 0$.

Sigmoidal (logistic) smoothing. With temperature $\gamma > 0$ let $\sigma_\gamma(z) = \big(1 + e^{-z/\gamma}\big)^{-1}$. The model

$$f_{\mathsf{DT},\gamma}(\boldsymbol{x}) \;=\; \prod_{i=1}^{D} \sigma_\gamma\big(\mathbf{w}_i^\top \boldsymbol{x} + b_i\big) \tag{3}$$

is infinitely differentiable; its transition width is $O(\gamma)$ and its spectrum decays polynomially in frequency, faster than the ramp yet slower than the Gaussian surrogate below.

Gaussian smoothing. Global diffusion is obtained by convolving the hard tree with an isotropic Gaussian kernel $G_\sigma(z) = (2\pi\sigma^2)^{-d/2} \exp(-\|z\|^2/(2\sigma^2))$:

$$f_\sigma(\boldsymbol{x}) \;=\; \int_{\mathbb{R}^d} f_{\mathsf{DT}}(\boldsymbol{y})\, G_\sigma(\boldsymbol{x} - \boldsymbol{y})\, d\boldsymbol{y}. \tag{4}$$

This surrogate is C^∞ with Fourier transform $\widehat{f_\sigma}(\xi) \;=\; e^{-\sigma^2\|\xi\|^2/2}\widehat{f_{\mathsf{DT}}}(\xi)$, implying exponential spectral decay. Unlike the ramp or sigmoid constructions—which preserve the separable, axis-aligned product structure—Gaussian convolution couples all coordinates, spreading the effect of each split over a neighbourhood of radius σ.

Each smoothing scheme recovers f_{DT} in the limit $\epsilon, \gamma, \sigma \to 0$, but exhibits markedly different regularity and spectral behaviour; these differences will be central to our subsequent analysis of their Radon total-variation norm.

Radon transform. For $f \in L^1(\mathbb{R}^d)$ the Radon transform is

$$\mathcal{R}f(\beta, t) := \int_{\{x:\, \beta^\top x = t\}} f(x)\, ds(x), \qquad (\beta, t) \in \mathbb{S}^{d-1} \times \mathbb{R},$$

where ds denotes the $(d-1)$-Lebesgue measure on the hyperplane. With the unitary Fourier convention $\widehat{f}(\xi) \;=\; (2\pi)^{-d/2} \int f(x) e^{-i\xi^\top x}\, dx$ the *Fourier–slice theorem* (see, e.g., Kak and Slaney (1988)) gives

$$\mathcal{R}f(\beta, t) \;=\; (2\pi)^{\frac{1-d}{2}} \int_{\mathbb{R}} e^{i\omega t}\, \widehat{f}(\omega\beta)\, d\omega. \tag{5}$$

Second-order Radon bounded-variation space. Following Ongie et al. (2020); Parhi and Nowak (2021) we define

$$\mathcal{R}\mathrm{BV}^2(\mathbb{R}^d) := \big\{ f \in L^{\infty,1}(\mathbb{R}^d) : \|f\|_{\mathcal{R}} < \infty \big\},$$

where

$$\|f\|_{\mathcal{R}} := c_d \left\| \partial_t^2 \Lambda^{d-1} \mathcal{R} f \right\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})}, \qquad c_d^{-1} = 2(2\pi)^{d-1}, \tag{6}$$

and $\Lambda^{d-1} = (-\partial_t^2)^{(d-1)/2}$ is the 1-D "ramp filter" operator.[1] The norm $\|\cdot\|_{\mathcal{R}}$ coincides with the minimum-width, infinite-neuron ReLU network norm introduced by Ongie et al. (2020), and measures the second-order total variation of $\mathcal{R}\{f\}$ across all projection directions. In this work we denote this seminorm by $\|\cdot\|_{\mathcal{R}}$.

Using the discussion above, it turns out the computation of $\|\cdot\|_{\mathcal{R}}$ in the one-dimensional setting can be simplified. This has been formally proven in Savarese et al. (2019) as follows:

Theorem 1 (Theorem 3.1 Savarese et al. (2019)) *For any function $f : \mathbb{R} \to \mathbb{R}$, we have:*

$$\mathcal{R}\mathrm{TV}^2(f) = \max\left(\int_{-\infty}^{\infty} |f''(x)|dx, |f'(\infty) + f'(-\infty)| \right) \leq \int_{-\infty}^{\infty} |f''(x)|dx + 2 \inf_x |f'(x)|$$

In higher dimension, we follow a three-step procedure to compute $\|\cdot\|_{\mathcal{R}}$.

3.1 A universal three-step recipe for $\|f\|_{\mathcal{R}}$

Eq. (5) and linearity yield the following computational pattern, which we employ in this work.

1. **Fourier transform.** Obtain (exactly or up to an explicit bound) the Fourier transform $\widehat{f}(\xi)$.
2. **One-dimensional Radon slice.**
   Substitute $\widehat{f}(\omega\beta)$ into (5) and differentiate once more in t:
   $$\partial_t^{d+1} \mathcal{R} f(\beta, t) = (2\pi)^{(1-d)/2} \int_{\mathbb{R}} (i\omega)^{d+1} e^{i\omega t}\, \widehat{f}(\omega\beta)\, d\omega.$$
3. **L^1-norm of the $(d+1)$st derivative.**
   Integrate the absolute value over $t \in \mathbb{R}$ and $\beta \in \mathbb{S}^{d-1}$, applying Fubini/Tonelli and any required bounds on $\widehat{f}$ to obtain $\|f\|_{\mathcal{R}}$.

Intuitively, Step 1 encodes geometric information in frequency space, Step 2 converts that information into directional line integrals, and Step 3 aggregates the variation of these integrals to yield the Radon BV norm. The recipe is agnostic to the specific form of f; it applies verbatim to the hard tree, ramp-smoothed, sigmoid-smoothed and Gaussian-smoothed models introduced earlier, differing only in the bounds used for $\widehat{f}$.

This formalises the procedure implicit in previous discussions: each $\|\cdot\|_{\mathcal{R}}$-norm computation reduces to a Fourier bound followed by a one-D integration in the projection variable t.

1. For odd d the fractional power is interpreted via Fourier multipliers; all derivatives are taken in the sense of tempered distributions.

4 Approximation of Hard-threshold Decision Trees via Shallow Networks

In Section 3 we introduced the Radon total-variation norm $\|\cdot\|_{\mathcal{R}}$ and its explicit form in Eq. (6). We now establish that this norm is *unbounded* for hard-threshold decision trees (Theorem 3). We start with the one–dimensional setting, where a decision tree reduces to a step function taking values in $\{0, 1\}$.

Consider the step function in single dimension denoted as $f_{\mathsf{step}} : \mathbb{R} \to \mathbb{R}$, defined as

$$f_{\mathsf{step}}(x) = \sum_{i=1}^{n} c_i \cdot 1\{x \in (z_i, z_{i+1})\}$$

for given set of scalars $-\infty < z_1 \leq z_2 \leq \ldots \leq z_N < \infty$.

Lemma 2 *$\mathcal{R}\mathrm{TV}^2(f_{\mathsf{step}})$ is unbounded.*

Proof [Proof outline] For $d = 1$, $\|f\|_{\mathcal{R}} = \int_{\mathbb{R}} |f''(x)|\, dx$. Each jump at $x = z_i$ yields $f'' = c_i\, \delta'_{z_i}$, where δ'_{z_i} is a *dipole distribution* ($\langle \delta'_{z_i}, \varphi \rangle = -\varphi'(z_i)$). Approximating δ'_{z_i} by the mollifier $\delta'_\varepsilon(x - z_i) = \varepsilon^{-2}\psi'((x - z_i)/\varepsilon)$ gives $\|\delta'_{z_i}\|_{L^1} = \lim_{\varepsilon \to 0} \varepsilon^{-2} \int |\psi'(u)|\, du = \infty$. Hence

$$\int_{\mathbb{R}} |f''(x)|\, dx = \sum_{i=1}^{n} |c_i|\, \|\delta'_{z_i}\|_{L^1} = \infty,$$

so the RTV diverges. ■

The divergence stems from the $(d+1)$-st derivative of an indicator, which contains derivatives of the Dirac delta distribution. These derivatives have infinite total-variation (equivalently, ℓ_1) norm, so no amount of averaging can regularise them. Because every axis-aligned decision tree contains a one–dimensional slice exhibiting the same pathology, the Radon-TV remains unbounded in any ambient dimension. The formal statement is given in Theorem 3; its proof is deferred to the supplemental materials.

Theorem 3 *Consider the decision tree $f_{\mathsf{DT}} : \mathbb{R}^d \to \mathbb{R}$ such that $f_{\mathsf{DT}}(\boldsymbol{x}) := 1\{\boldsymbol{x} \in A\}$ for an axes aligned compact subset $A \subset \mathbb{R}^d$. Then, $\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT}})$ of the decision tree as defined is unbounded.*

Proof [Proof outline] First observe that the Fourier transform of the indicator over the axis-aligned box A is

$$\hat{f}_{\mathsf{DT}}(\boldsymbol{\xi}) \;=\; (2\pi)^{-d/2} \int_A e^{-\mathrm{i}\boldsymbol{\xi}^\top \boldsymbol{x}}\, d\boldsymbol{x}.$$

Consequently, for odd dimension d the Radon-TV norm becomes

$$\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT}}) = c_d (2\pi)^{-(d-1)/2} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big| \int_A \delta^{(d+1)}\big(t - \boldsymbol{\beta}^\top \boldsymbol{x}\big)\, d\boldsymbol{x} \Big|\, dt\, d\boldsymbol{\beta}.$$

Fix the direction $\boldsymbol{\beta}_0 = \mathbf{e}_1 \in \mathbb{S}^{d-1}$ and let $\mathcal{B}(\boldsymbol{\beta}_0, \epsilon) \subseteq \mathbb{S}^{d-1}$ be the spherical cap of radius $\epsilon > 0$. For each $\boldsymbol{\beta} \in \mathbb{S}^{d-1}$ define

$$g_{\boldsymbol{\beta}}(u) := \int_A \delta\big(u - \boldsymbol{\beta}^\top \boldsymbol{x}\big)\, d\boldsymbol{x}.$$

By the co-area formula (Mattila, 1995), this can be rewritten as

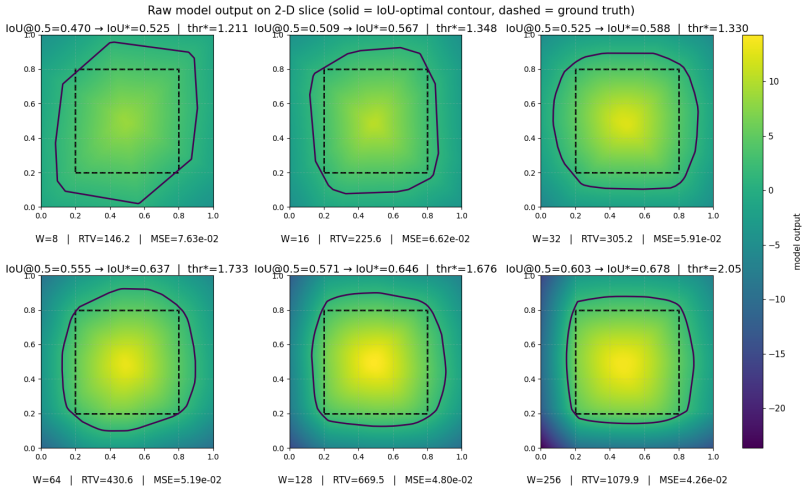


Figure 2: Raw network output on a 2-D slice of $[0,1]^5$ for single-hidden-layer ReLU nets of increasing width W. The target set is the axis-aligned box $[0.2, 0.8]^5$. For visualization we fix three coordinates at 0.5 and vary $(x_0, x_1) \in [0,1]^2$. The heatmap shows the model score $f_\theta(x)$ (pre-sigmoid *logit*); the **solid** curve is the decision boundary $\{x : f_\theta(x) = \tau^\star\}$ at the *IoU-optimal* threshold $\tau^\star$ chosen on a validation split; the **dashed** rectangle is the ground-truth boundary on this slice. Panel headers report IoU@0.5 $\rightarrow$ IoU* and $\tau^\star$; panel footers report W, the squared-weight complexity $\mathcal{R}\mathrm{TV}^2 = \frac{1}{2}\big(\sum_i \|w_i\|_2^2 + \sum_i a_i^2\big)$ (biases excluded), and test MSE (using $\sigma(f_\theta)$ for BCE training). Intersection-over-Union (IoU) is computed on the full 5-D test set as $\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP+FP+FN}}$, i.e., the measure of predicted-positive $\cap$ true-positive divided by that of their union; it ignores TN so it reflects *shape overlap*. As W increases, the learned boundary aligns better with the dashed box and IoU rises, while RTV also grows, illustrating a fit–complexity trade-off and the benefit of tuning the threshold vs. using a fixed 0.5.

$$g_{\boldsymbol{\beta}}(u) = \int_{A \cap \{\boldsymbol{x}:\boldsymbol{\beta}^\top \boldsymbol{x} = u\}} \frac{1}{\|\boldsymbol{\beta}\|}\, d\sigma(\boldsymbol{x}) \;=\; \mathrm{Vol}_{d-1}\big(A \cap \{\boldsymbol{x} : \boldsymbol{\beta}^\top \boldsymbol{x} = u\}\big),$$

where $d\sigma$ denotes the $(d-1)$-dimensional Hausdorff measure on the hyperplane $\{\boldsymbol{x} : \boldsymbol{\beta}^\top \boldsymbol{x} = u\}$.

For $\boldsymbol{\beta}_0 = \mathbf{e}_1$ this is the step function

$$g_{\mathbf{e}_1}(u) = \begin{cases} 0, & u \notin [a_1, b_1], \\ \mathrm{Vol}_{d-1}\big([a_2, b_2] \times \cdots \times [a_d, b_d]\big), & u \in (a_1, b_1). \end{cases}$$

Hence $g_{\mathbf{e}_1}$ has sharp jumps at $u = a_1$ and $u = b_1$. Perturbing to $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \Delta$ with $\|\Delta\| \le \epsilon$ smooths these jumps over an $O(\epsilon)$ interval, but the slope grows rapidly. In particular, for sufficiently small $\epsilon > 0$ there exists $u_\epsilon \approx a_1$ such that

$$\big|g^{(d+1)}_{\boldsymbol{\beta}_0+\Delta}(u_\epsilon)\big| \;\approx\; \frac{C}{\epsilon^{d+1}}$$

for some constant $C > 0$. Integrating over a window of width ϵ yields

$$\int_{\mathbb{R}} \big|g^{(d+1)}_{\boldsymbol{\beta}_0+\Delta}(u)\big|\, du \;\ge\; \frac{C'}{\epsilon^d}, \;\text{ with } C' > 0.$$

Now integrate this lower bound over the cap $\mathcal{B}(\boldsymbol{\beta}_0, \epsilon)$, whose surface measure scales as ϵ^{d-1}:

$$\int_{\mathcal{B}(\boldsymbol{\beta}_0,\epsilon)} \int_{\mathbb{R}} \left| g_{\boldsymbol{\beta}}^{(d+1)}(u) \right| du \, d\boldsymbol{\beta} \;\geq\; \epsilon^{d-1} \, \frac{C'}{\epsilon^d} \;=\; \frac{C'}{\epsilon}.$$

Because this bound holds for every sufficiently small ϵ, letting $\epsilon \to 0$ forces

$$\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT}}) = \infty.$$

The same argument applies to any canonical direction, completing the proof for general axis-aligned decision trees. ■

Remark 4 (Implication for shallow networks) *Any single-hidden-layer ReLU network whose weights are chosen so that the induced function has bounded Radon-TV cannot approximate a hard-threshold decision tree to arbitrary accuracy. Additional depth or unbounded weight growth is necessary.*

5 Approximation of Smoothed Decision Trees via Shallow Networks

Section 4 showed that the Radon total–variation (RTV) norm is infinite for hard-threshold decision trees. At first sight this is puzzling: the universal approximation theorem guarantees that even a *shallow* neural network can approximate any continuous function to arbitrary accuracy. The catch is the discontinuity—no finite RTV ball, regardless of width, can capture a step. We therefore ask whether *smoothing* the tree reduces the norm.

5.1 Piecewise-linear (ramp) smoothing

Equation (2) replaces each hard split by a centred ramp, producing a continuous piecewise-linear function. The behaviour differs sharply between one- and higher-dimensional domains.

The one–dimensional case. For $d = 1$ a depth-D ramp tree has exactly $2D$ kink points

$$s_1 < s_2 < \cdots < s_{2D}, \qquad s_{2i-1/2} \;=\; -\frac{b_i}{w_i} \;\mp\; \frac{\epsilon}{2w_i}, \quad i = 1, \ldots, D,$$

and is affine on each open interval (s_k, s_{k+1}) with slope m_k. Writing $c_0 := f_{\mathsf{DT},\epsilon}(x_0)$ for some $x_0 < s_1$ and $c_k := m_k - m_{k-1}$ for $k = 1, \ldots, 2D$, we obtain

$$f_{\mathsf{DT},\epsilon}(x) \;=\; c_0 + \sum_{k=1}^{2D} c_k \, [x - s_k]_+.$$

Hence a single-hidden-layer ReLU network with $2D$ units (unit input weights, biases $-s_k$, and output weights c_k) represents the ramp tree *exactly* on $\mathbb{R}$; the function lies in $\mathcal{R}\mathrm{BV}^2(\mathbb{R})$.

The multi-dimensional case. When $d > 1$ the situation changes. If $D \leq 2$ and the normals $\boldsymbol{w}_1, \boldsymbol{w}_2$ are parallel, the RTV remains finite. For $D \geq 3$ the ramp tree generally has infinite Radon norm. The following result from Ongie et al. (2020) applies.

Proposition 5 (Proposition 5-(a) of Ongie et al., 2020) *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous piecewise linear function with compact support. If at least one of the boundary normals is not parallel with every other boundary normal. Then f has infinite $\mathcal{R}$-norm.*

In a ramp-smoothed tree the hyperplanes $\{\boldsymbol{x} : (\boldsymbol{w}_i^\top \boldsymbol{x} + b_i)/\epsilon + \frac{1}{2} = 0\}$ intersect the planes $\{(\mathbf{w}, b) = (\mathbf{0}, 0)\}$ and $\{(\mathbf{w}, b) = (\mathbf{0}, 1)\}$ along $(d-1)$-dimensional faces, so condition (a) is met whenever the ambient dimension satisfies $d > 1$. Moreover, constructing a continuous approximation to a hard threshold requires at least four ramps ($D \geq 4$), ensuring the hypothesis of Proposition 5. Consequently, for $d > 1$ and realistic depths the Radon norm of a piecewise-linear smoothed tree is still infinite.

5.2 Smoothed Decision Trees

We now investigate two differentiable smoothing schemes that exhibit markedly different spectral decay: the *sigmoidal* and the *Gaussian* smoothings. For the sigmoidal case we show that the Radon norm remains infinite (Theorem 6), whereas Gaussian smoothing yields a finite norm whose magnitude depends explicitly on the ambient dimension (Theorem 7).[2]

Sigmoidal smoothing. For any $\gamma > 0$ and shift $b \in \mathbb{R}$ the Fourier transform of the scaled logistic

$$\sigma_\gamma(z) \;=\; \tfrac{1}{1+e^{-z/\gamma}} \;=\; \tfrac{1}{2} + \tfrac{1}{2}\tanh\big(z/(2\gamma)\big)$$

is

$$\widehat{\sigma_\gamma(\,\cdot + b)}(\omega) \;=\; e^{-\mathrm{i}\omega b}\Big[\frac{\pi}{2}\,\delta(\omega) + \frac{\mathrm{i}\,\gamma\pi}{\sinh(\pi\gamma\omega)}\Big], \qquad \omega \in \mathbb{R}. \tag{7}$$

Using Eq. (7), the Fourier transform of the sigmoidal D-split tree

$$f_{\mathsf{DT},\gamma}(\boldsymbol{x}) \;:=\; \prod_{i=1}^{D} \sigma_\gamma\big(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i\big), \qquad \{\boldsymbol{w}_i\}_{i=1}^D \subset \mathbb{R}^d,$$

with orthonormal normals $\boldsymbol{w}_i$, is

$$\widehat{f_{\mathsf{DT},\gamma}}(\boldsymbol{\xi}) \;=\; (2\pi)^{-(d-D)/2}\,\delta\big(P_{\mathcal{S}^\perp}\boldsymbol{\xi}\big)\prod_{i=1}^{D} e^{-\mathrm{i}\,b_i\eta_i}\Big[\tfrac{\pi}{2}\,\delta(\eta_i) + \tfrac{\mathrm{i}\,\gamma\pi}{\sinh(\pi\gamma\eta_i)}\Big], \quad \eta_i = \boldsymbol{w}_i^\top\boldsymbol{\xi},$$

where $\mathcal{S} = \operatorname{span}\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D\}$. Substituting this expression into the Radon norm definition (6) yields the following theorem. We defer the proof to the supplemental materials.

Theorem 6 *Fix a depth $D \geq 1$ and temperature $\gamma > 0$, and let*

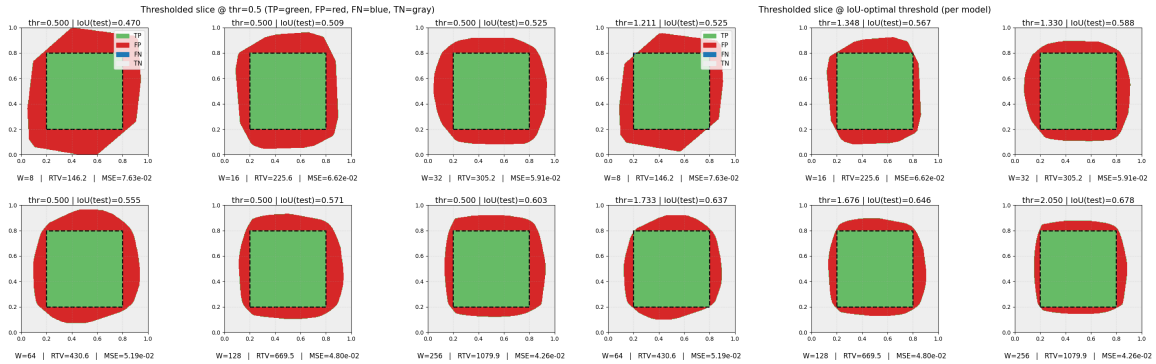
$$f_{\mathsf{DT},\gamma}(\boldsymbol{x}) \;=\; \prod_{i=1}^{D} \sigma_\gamma\big(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i\big), \qquad \{\boldsymbol{w}_i\}_{i=1}^D \text{ orthonormal and distinct}.$$

Then

$$\|f_{\mathsf{DT},\gamma}\|_{\mathcal{R}} = \frac{c_d}{\sqrt{2\pi}}\,(2\pi)^{-\frac{d-D}{2}}\,(\gamma\pi)^D \int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}} |\omega|^{d+1}\,\mathbf{1}_{\{\boldsymbol{\beta}\in\mathcal{S}\}}\prod_{i=1}^{D}\frac{1}{\big|\sinh\big(\pi\gamma\lambda_i(\boldsymbol{\beta})\,\omega\big)\big|}\,d\omega\,d\sigma(\boldsymbol{\beta}),$$

where $\lambda_i(\boldsymbol{\beta}) = \boldsymbol{w}_i^\top\boldsymbol{\beta}$. In particular, for every $D \geq 2$ the Radon norm diverges: $\|f_{\mathsf{DT},\gamma}\|_{\mathcal{R}} = \infty$.

2. For one–dimensional signals certain smoothings do produce a bounded Radon norm; the argument is analogous to the $d = 1$ ramp analysis and is omitted for brevity.



(a) Threshold at 0.5 (TP=green, FP=red, FN=blue, TN=gray).

(b) Threshold at tuned (IoU-optimal) level.

Figure 3: Effect of thresholding on the same 2-D slice. Each subplot shows the binary mask of the model's positive region (green = TP, red = FP, blue = FN, gray = TN) with the ground-truth box outlined (dashed). Top row: fixed threshold 0.5; bottom row: the IoU-optimal threshold $\tau^\star$ chosen on a validation split. Numbers above each subplot give the threshold used and the resulting test IoU; numbers below give W, $\mathcal{R}\mathrm{TV} = \frac{1}{2}\big(\sum_i \|w_i\|_2^2 + \sum_i a_i^2\big)$, and test MSE. IoU is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$ on the full 5-D test set. Tuning the threshold typically reduces both FP and FN along the box boundary, yielding higher IoU without changing the underlying network.

Note that for $D = 1$, as noted in the case of linear ramp smoothing, it is straight-forward to note that $\|f_{\mathsf{DT},\gamma}\|_{\mathcal{R}}$ is bounded.

Gaussian smoothing. Convolving the hard tree with an isotropic Gaussian rapidly suppresses high-frequency components: in the Fourier domain $\widehat{f_\sigma}(\boldsymbol{\xi}) = \widehat{f_{\mathsf{DT}}}(\boldsymbol{\xi})\, e^{-\sigma^2\|\boldsymbol{\xi}\|^2/2}$, so every factor of $\|\boldsymbol{\xi}\|$ in the Radon norm integrand is offset by Gaussian decay. We state the main theorem that quantifies the gain, with the proof deferred to the supplemental materials.

Theorem 7 *Fix any bandwidth $\sigma > 0$. For the Gaussian smoothed tree*

$$f_\sigma(\boldsymbol{x}) = \int_{\mathbb{R}^d} f_{DT}(\boldsymbol{y})\, G_\sigma(\boldsymbol{x}-\boldsymbol{y})\, d\boldsymbol{y}, \qquad G_\sigma(z) = (2\pi\sigma^2)^{-d/2} e^{-\|z\|^2/(2\sigma^2)},$$

the Radon-total-variation satisfies $\|f_\sigma\|_{\mathcal{R}} \;\le\; C\, d^{1/2} \left(\frac{\sqrt{2}\, e}{\sigma}\right)^d Vol(A)$, for $sd \ge 1$, with a universal constant $C \le 1.2$.

6 Classification Problem is Easy

In the previous sections we showed that exact representation of the hard tree hits a bottleneck with shallow networks which could be thought of a regression task. Here, we show that the problem is easier if we only care about *classification* after thresholding a smooth score, rather than exact function approximation of the hard tree itself. We separate two goals: (i) exact recovery of the decision set A after thresholding a smooth score, and (ii) control of calibration ($\mathrm{L}^1(P)$ closeness to $\mathbf{1}_A$) and complexity ($\mathcal{R}\mathrm{TV}$). Throughout, A is a single axis-aligned box in $\mathbb{R}^d$ (the results can be extended to a finite union as well).

Construction. Fix $\lambda \geq 1$ and set $\varepsilon = c_0/\lambda$ with $c_0 > 0$. Let $H \in C^\infty(\mathbb{R})$ be nondecreasing with $H(s) = 0$ for $s \leq 0$, $H(s) = 1$ for $s \geq 1$, and $H^{(m)}(0) = H^{(m)}(1) = 0$ for $1 \leq m \leq d+1$. Define

$$h_\varepsilon(t) := H\big((t+\varepsilon)/\varepsilon\big), \qquad \vartheta_{\lambda,\varepsilon}(t) := (1 - h_\varepsilon(t))\, e^{\lambda t} + h_\varepsilon(t).$$

For a box $B = \prod_{j=1}^d [\ell_j, u_j]$, set

$$S_B(x) \;:=\; \prod_{j=1}^d \vartheta_{\lambda,\varepsilon}(u_j - x_j)\, \vartheta_{\lambda,\varepsilon}(x_j - \ell_j) \;\in [0,1].$$

Exact thresholding and calibration. We quantify classification (exact recovery at a single cutoff) and $\mathrm{L}^1(P)$ calibration under a standard tube–mass condition around ∂B.

Assumption 1 (Tube–mass near ∂B) *There exist $C > 0$ and $\beta > 0$ such that, for all sufficiently small $t > 0$ and $X \sim P$,*

$$\mathsf{T}_B(t) := P\{\mathrm{dist}(X, \partial B) \leq t\} \;\leq\; C\, t^\beta.$$

Lemma 8 (Single box: exact thresholding and distributional closeness) *Let S_B be defined as above.Then, under Assumption 1 for all $\lambda \geq 1$:*

$$\{x :\; S_B(x) \geq 1\} = B, \qquad \mathbb{E}\big[\,|S_B(X) - \mathbf{1}_B(X)|\,\big] \;\leq\; C_{d,\beta,c_0}\, C\, \lambda^{-\beta}.$$

Proof deferred to Appendix E.

So, we have established that if we model a problem for classification with an axis-aligned box decision boundary, then there exists a smooth score S_B that thresholds exactly to the box at cutoff 1, and its $\mathcal{L}^1(P)$ distance to the hard indicator $\mathbf{1}_B$ decays as $O(\lambda^{-\beta})$ under the tube-mass condition on the data distribution. But beyond that, this score is further amenable to complexity control in terms of its Radon total-variation norm, as we show next. Thus, this score is representable by a finite-width ReLU network with bounded norm.

Theorem 9 (RTV upper bound for a single box) *For all $\lambda \geq 1$ and $\varepsilon = c_0/\lambda$,*

$$\|S_B\|_{\mathcal{R}} \;\leq\; C_d \sum_{r=1}^d \lambda^{\,d+1-r}\, \mathcal{H}^{d-r}\big(\Sigma_{d-r}(B)\big),$$

where C_d depends only on d (and the choice of H in $\vartheta_{\lambda,\varepsilon}$) and is independent of λ.

where $\Sigma_{d-r}(B)$ denotes the union of $(d-r)$–dimensional faces of B, and $\mathcal{H}^{d-r}(\Sigma_{d-r}(B))$ denotes the $(d-r)$–dimensional Hausdorff measure.

We defer the proof to Appendix E.

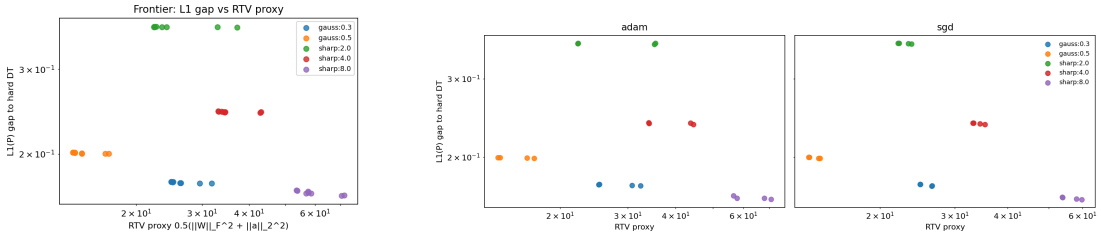


Figure 4: **Learning a calibrated score.** Empirical score–smoothness frontier for shallow ReLU nets trained on smoothed tree targets. We plot *raw scores* (no thresholding): x–axis = RTV proxy $\frac{1}{2}(\|W\|_F^2 + \|a\|_2^2)$, y–axis = $\mathrm{L}^1(P)$ gap $E[|s(x) - \mathbf{1}_A(x)|]$ (uniform P; Monte Carlo). Colors indicate the target family ($\texttt{gauss}:\sigma$, $\texttt{sharp}:k$). Increasing sharpness (larger k, smaller σ) raises RTV and lowers the L^1 gap, producing a clear frontier. **Left:** all runs (widths 64/128; wd 0 or 10^{--4}; Adam/SGD). **Right:** split by optimizer–both trace similar frontiers but land at different points (i.e., different RTV–accuracy trade-offs).

7 Conclusion and Future Works

We studied how *shallow* ReLU networks approximate the decision regions of axis-aligned trees through the lens of the Radon total-variation (RTV) seminorm. Our analysis separates two tasks that are often conflated: (i) *classification via thresholding*–which is easy–and (ii) *score learning / calibration*–which exhibits a quantitative trade-off. The hard tree $\mathbf{1}_A$ has infinite RTV, several naïve smoothings inherit this, and our "sharp barrier" scores s_k provide a constructive alternative: they admit *finite* RTV with $\|s_k\|_{\mathcal{R}} \lesssim k\,\mathcal{H}^{d-1}(\partial A)$, have an *exact* threshold that recovers A, and achieve $\mathrm{L}^1(P)$ error $\lesssim k^{-\beta}$ under a mild tube-mass condition. Empirically, different optimizers populate different points along this frontier, but the thresholded classifier is already near-perfect across a wide RTV range.

Future Directions. (i) Beyond two layers, RTV is typically infinite; sharper, depth-aware complexity measures are needed to probe whether an accuracy–interpretability gap persists for deep nets. (ii) Calibrated, low-RTV scores align gradients with tree features; understanding when optimization biases training toward such aligned solutions is an open question. (iii) Extending the analysis to oblique trees and data-dependent norms, and coupling training with explicit RTV-style regularization, are promising directions.

References

Pranjal Atrey, Michael P. Brundage, Min Wu, and Sanghamitra Dutta. Demystifying the accuracy-interpretability trade-off: A case study of inferring ratings from reviews, 2025. URL https://arxiv.org/abs/2503.07914.

Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks, 2017. URL https://arxiv.org/abs/1706.08498.

George V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989. URL https://api.semanticscholar.org/CorpusID:3958369.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. URL https://arxiv.org/abs/1702.08608.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks, 2016. URL https://arxiv.org/abs/1512.03965.

Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree, 2017. URL https://arxiv.org/abs/1711.09784.

Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90003-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900038.

I.M. Gel'fand and G.E. Shilov. Chapter i - definition and simplest properties of generalized functions. In I.M. Gel'fand and G.E. Shilov, editors, *Properties and Operations*, pages 1–151. Academic Press, 1964. ISBN 978-1-4832-2976-8. doi: https://doi.org/10.1016/B978-1-4832-2976-8.50007-6. URL https://www.sciencedirect.com/science/article/pii/B9781483229768500076.

Hussein Hazimeh, Natalia Ponomareva, Petros Mol, Zhenyu Tan, and Rahul Mazumder. The tree ensemble layer: Differentiability meets conditional computation, 2020. URL https://arxiv.org/abs/2002.07772.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572.

A C Kak and M Slaney. *Principles of computerized tomographic imaging*. IEEE Service Center,Piscataway, NJ, 01 1988. URL https://www.osti.gov/biblio/5813672.

Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, and Samuel Rota Bulo. Deep neural decision forests. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Alessandro Lovo, Amaury Lancelin, Corentin Herbert, and Freddy Bouchet. Tackling the accuracy-interpretability trade-off in a hierarchy of machine learning models for the prediction of extreme heatwaves, 2025. URL https://arxiv.org/abs/2410.00984.

Tong Mao, Jonathan W. Siegel, and Jinchao Xu. Approximation rates for shallow relu^k neural networks on sobolev spaces via the radon transform, 2024. URL https://arxiv.org/abs/2408.10996.

Sascha Marton, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. GRANDE: Gradient-based decision tree ensembles for tabular data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XEFWBxi075.

Pertti Mattila. *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

H. N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8(1):164–177, 1996. doi: 10.1162/neco.1996.8.1.164.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks, 2015. URL https://arxiv.org/abs/1506.02617.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1lNPxHKDH.

Subrat Prasad Panda, Blaise Genest, Arvind Easwaran, and Ponnuthurai Nagaratnam Suganthan. *Vanilla Gradient Descent for Oblique Decision Trees*. IOS Press, October 2024. ISBN 9781643685489. doi: 10.3233/faia240607. URL http://dx.doi.org/10.3233/FAIA240607.

Rahul Parhi and Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021. URL https://jmlr.org/papers/v22/20-583.html.

Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8: 143–195, 1999. doi: 10.1017/S0962492900002919.

Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data, 2019. URL https://arxiv.org/abs/1909.06312.

Bateman Manuscript Project, H. Bateman, A. Erdélyi, and United States. Office of Naval Research. *Tables of Integral Transforms: Based, in Part, on Notes Left by Harry Bateman, and Compiled by the Staff of the Bateman Manuscript Project. [A. Erdélyi, Editor. W. Magnus, F. Oberhettinger, F. G. Tricomi, Research Associates]*. Number v. 1 in California Institute of Technology: Bateman Manuscript Project. McGraw-Hill, 1954. URL https://books.google.com/books?id=OLfZAAAAMAAJ.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019. URL https://arxiv.org/abs/1811.10154.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2667–2690. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/savarese19a.html.

Ryutaro Tanno, Kai Arulkumaran, Daniel Alexander, Antonio Criminisi, and Aditya Nori. Adaptive neural trees. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6166–6175. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tanno19a.html.

Matus Telgarsky. benefits of depth in neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1517–1539, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/telgarsky16.html.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/yarotsky18a.html.

Appendix A. Is Interpretability at Odds with Accuracy? Inapproximability of Decision Trees by Shallow Networks: Supplementary Materials

Appendix B. Hard Threshold Decision Trees

In this appendix, we provide the proof of the main results as presented in Section 4.

B.1 $\mathcal{R}\mathrm{TV}^2$ of 1-D step functions

We defined a step function in single dimension as $f_{\mathsf{step}} : \mathbb{R} \to \mathbb{R}$ where

$$f_{\mathsf{step}}(x) = \sum_{i=1}^{n} c_i \cdot 1\{x \in (z_i, z_{i+1})\}$$

for given set of scalars $-\infty < z_1 \leq z_2 \leq \ldots \leq z_N < \infty$

We restate the claim of unboundedness of $\mathcal{R}\mathrm{TV}^2(f_{\mathsf{step}})$ with the proof below it.

Lemma 10 *$\mathcal{R}\mathrm{TV}^2(f_{\mathsf{step}})$ is unbounded.*

Proof Using Theorem 3.1 (Savarese et al., 2019), we note that for a choice of small enough $\epsilon > 0$

$$\begin{aligned}
\mathcal{R}\mathrm{TV}^2(f_{\mathsf{step}}) &= \int_{-\infty}^{\infty} |f''_{\mathsf{step}}(x)| dx \\
&= \int_{-\infty}^{\infty} \left| \left(\sum_{i=1}^{n} c_i \cdot \Delta_{i,i+1} \delta_{z_i}(x) \right)' \right| dx \\
&= \int_{-\infty}^{\infty} \left| \sum_{i=1}^{n} c_i \cdot \Delta_{i,i+1} \delta'_{z_i}(x) \right| dx \\
&= \sum_{i=1}^{n} \int_{z_i-\epsilon}^{z_i+\epsilon} \left| c_i \cdot \Delta_{i,i+1} \delta'_{z_i}(x) \right| dx \\
&= \sum_{i=1}^{n} c_i \cdot \Delta_{i,i+1} \int_{z_i-\epsilon}^{z_i+\epsilon} \left| \delta'_{z_i}(x) \right| dx \to \infty
\end{aligned}$$

where in the last equation we note that the δ' is a dipole distribution whose ℓ_1 norm is unbounded. ■

B.2 $\mathcal{R}\mathrm{TV}^2$ of high-dimensional step functions

In this appendix, we provide the proof of Theorem 3.

Throughout we adopt the *unitary* Fourier convention

$$\widehat{g}(\boldsymbol{\xi}) \;:=\; (2\pi)^{-d/2} \int_{\mathbb{R}^d} g(x)\, e^{-\mathrm{i}\,\boldsymbol{\xi}^\top \boldsymbol{x}}\, d\boldsymbol{x}, \qquad \boldsymbol{\xi} \in \mathbb{R}^d. \tag{8}$$

All computations are in the sense of **tempered distributions** (Schwartz space dual); every integral we write down exists in that sense.

Thus, for a decision tree as defined in Section 3:

$$\begin{aligned}
\hat{f}_{\mathsf{DT}}(\boldsymbol{\xi}) &:= (2\pi)^{-d/2} \int f_{\mathsf{DT}}(\boldsymbol{x}) e^{-\mathrm{i}\xi^\top \boldsymbol{x}} d\boldsymbol{x} \\
&= (2\pi)^{-d/2} \int 1\{\boldsymbol{x} \in A\}\, e^{-\mathrm{i}\boldsymbol{\xi}^\top \boldsymbol{x}} d\boldsymbol{x} \\
&= (2\pi)^{-d/2} \int_A e^{-\mathrm{i}\xi^\top \boldsymbol{x}} d\boldsymbol{x}
\end{aligned}$$

Evaluating $\hat{f}_{\mathsf{DT}}$ at $\boldsymbol{\omega} = \omega\boldsymbol{\beta}$, we get

$$\hat{f}_{\mathsf{DT}}(\omega\boldsymbol{\beta}) = (2\pi)^{-d/2} \int_A e^{-\mathrm{i}\omega\boldsymbol{\beta}^\top \boldsymbol{x}} d\boldsymbol{x} \tag{9}$$

Using the Fourier slice theorem (Kak and Slaney, 1988) we have

$$\mathcal{F}_1\{\mathcal{R}\{f_{\mathsf{DT}}\}(\boldsymbol{\beta}, \cdot)\}(\omega) = \hat{f}_{\mathsf{DT}}(\omega\boldsymbol{\beta}).$$

This gives the following Radon transform of f_{DT}:

$$\begin{aligned}\mathcal{R}\{f_{\mathsf{DT}}\}(\boldsymbol{\beta}, t) &= (2\pi)^{-1/2} \int_{\mathbb{R}} e^{\mathrm{i}\omega t} \hat{f}_{\mathsf{DT}}(\omega\boldsymbol{\beta}) d\omega \\ &= (2\pi)^{-(d+1)/2} \int_{\mathbb{R}} e^{\mathrm{i}\omega t} \int_A e^{-\omega \mathrm{i}\boldsymbol{\beta}^\top \boldsymbol{x}} d\boldsymbol{x}\, d\omega \\ &= (2\pi)^{-(d+1)/2} \int_A \int_{\mathbb{R}} e^{\mathrm{i}\omega(t-\boldsymbol{\beta}^\top \boldsymbol{x})} d\omega\, d\boldsymbol{x} \\ &= (2\pi)^{-(d+1)/2} \int_A \delta(t - \boldsymbol{\beta}^\top \boldsymbol{x}) d\boldsymbol{x}\end{aligned}$$

Now, we compute the $d+1$-derivative of $\mathcal{R}\{f\}(\boldsymbol{\beta}, t)$ with respect to t (the integral is defined in the sense of a tempered distribution and follows the convention discussed in Gel'fand and Shilov (1964))

$$\partial_t \mathcal{R}\{f_{\mathsf{DT}}\}(\boldsymbol{\beta}, t) = (2\pi)^{-(d+1)/2} \int_A \delta'(t - \boldsymbol{\beta}^\top \boldsymbol{x})\, d\boldsymbol{x}$$

Similarly $(d+1)$th derivative in t is

$$\partial_t^{d+1} \mathcal{R}\{f_{\mathsf{DT}}\}(\boldsymbol{\beta}, t) = (2\pi)^{-(d+1)/2} \int_A \delta^{(d+1)}(t - \boldsymbol{\beta}^\top \boldsymbol{x})\, d\boldsymbol{x}$$

If d is odd, then the second-order Radon domain total variation is the L^1-norm of $(d+1)$ derivatives in t of this quantity (see Equation.(28) in Parhi and Nowak (2021)). That is

$$\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT}}) = c_d (2\pi)^{-(d-1)/2} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left| \int_A \delta^{(d+1)}(t - \boldsymbol{\beta}^\top \boldsymbol{x})\, d\boldsymbol{x} \right| dt\, d\boldsymbol{\beta}$$

Lets define for any $\boldsymbol{\beta} \in \mathbb{S}^{d-1}$

$$g_{\boldsymbol{\beta}}(u) := \int_A \delta(u - \boldsymbol{\beta}^\top \boldsymbol{x})\, d\boldsymbol{x} \tag{10}$$

where we can write

$$g_{\boldsymbol{\beta}}^{(k)}(u) = \int_A \delta^{(k)}(u - \boldsymbol{\beta}^\top \boldsymbol{x})\, d\boldsymbol{x}$$

for any $k > 0$.

Now, using the co-area formula (Mattila, 1995), noting that $||\boldsymbol{\beta}|| = 1$, we can rewrite Eq. (10) as

$$\int_A \delta(u - \boldsymbol{\beta}^\top \boldsymbol{x})\, d\boldsymbol{x} = \int_{\boldsymbol{\beta}^\top \boldsymbol{x} = u} 1\{\boldsymbol{x} \in A\}\, d\sigma(\boldsymbol{x})$$

where $d\sigma$ denotes the $(d-1)$-dimensional Hausdorff measure on the hyperplane $\{\boldsymbol{x} : \boldsymbol{\beta}^\top \boldsymbol{x} = u\}$

Now, we will show there exists, $\boldsymbol{\beta_0} \in \mathbb{S}^{d-1}$, and scalar $\epsilon > 0$ such that for

$$\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT}}) \geq \int_{\mathcal{B}(\boldsymbol{\beta_0},\epsilon)} \int_{\mathbb{R}} |g_{\boldsymbol{\beta}}^{(d+1)}(t)|\, dt\, d\boldsymbol{\beta} \to \infty \tag{11}$$

where $\mathcal{B}_2(\boldsymbol{\beta_0}, \epsilon) := \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \leq \epsilon\}$.

Without loss of generality, assume that A is axes-aligned to eigendirections of $\mathbb{R}^d$— $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d$. Thus, consider the case where $\boldsymbol{\beta_0} := \mathbf{e}_1$.

Now,

$$g_{\boldsymbol{\beta}_0+\Delta}(u) = \begin{cases} 0, & \text{if } u \in (a_1, b_1)^c \\ Vol\left(\left\{(\boldsymbol{\beta}_0 + \Delta)^\top \boldsymbol{x} = u\right\} \cap A\right) & \text{if } u \in (a_1, b_1) \end{cases}$$

Note, if $\Delta = \mathbf{0}$, then $g_{\boldsymbol{\beta}_0}$ has a sharp discontinuity at $u = a_1$. But if $\Delta \neq \mathbf{0}$ and $||\boldsymbol{\beta}_0 - \Delta|| \leq \epsilon$, $g_{\boldsymbol{\beta}_0+\Delta}$ varies over the real line smoothly. But we can control the jump around $(a_1 - \epsilon, a_1 + \epsilon)$. Note that,

$$\lim_{\epsilon \to 0} g_{\boldsymbol{\beta}_0+\Delta} = g_{\boldsymbol{\beta}_0}$$

Fix a Gaussian mollifier $\eta \in C_c^\infty(\mathbb{R})$, $\eta(u) := \pi^{-1/2} e^{-u^2}$ and put $\eta_\epsilon(u) = \epsilon^{-1}\eta(u/\epsilon)$, $g_{\boldsymbol{\beta},\epsilon} = g_{\boldsymbol{\beta}} * \eta_\epsilon$. If $h(u) = H\mathbf{1}_{\{u \geq 0\}}$ then $\max_u |(h * \eta_\epsilon)^{(k)}(u)| = H\, \epsilon^{-(k+1)} \max_s |\eta^{(k)}(s)|$.

Every slice with $\boldsymbol{\beta} \in \mathcal{B}(\boldsymbol{\beta}_0, \epsilon)$ still contains a jump of height at least $H/2$, where $H = \prod_{j=2}^d (b_j - a_j)$. With $k = d + 1$ this gives

$$\max_t |g_{\boldsymbol{\beta},\epsilon}^{(d+1)}(t)| \geq C\, \epsilon^{-(d+2)}, \qquad \int_{\mathbb{R}} |g_{\boldsymbol{\beta},\epsilon}^{(d+1)}(t)|\, dt \geq C\, \epsilon^{-(d+1)}.$$

Hence

$$\int_{\mathcal{B}(\boldsymbol{\beta}_0,\epsilon)} \int_{\mathbb{R}} |g_{\boldsymbol{\beta},\epsilon}^{(d+1)}(t)|\, dt\, d\boldsymbol{\beta} \;\geq\; C'\, \epsilon^{\,d-1}\, \epsilon^{-(d+1)} = \frac{C'}{\epsilon^2} \xrightarrow{\epsilon \to 0} \infty.$$

But note that as ϵ tends to 0, the mollified function $g_{\beta,\epsilon}^{(d+1)}$ tends to $g_\beta^{(d+1)}$ in distribution. But then this implies that over the convex slope $\mathcal{B}_2(\boldsymbol{\beta}_0, \epsilon)$

$$\int_{\mathcal{B}(\boldsymbol{\beta}_0)} \int_{\mathbb{R}} |g_{\boldsymbol{\beta},\epsilon}^{(d+1)}(t)|\, dt\, d\boldsymbol{\beta} \to \infty$$

as $\epsilon \to 0$. Hence, using the bound in Eq. (11), $\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT}})$ is unbounded.

Appendix C. Sigmoidal Smoothing

In this appendix, we will provide the proof of Theorem 6 as stated in Section 5 on the approximation of Sigmoid smooth decision trees.

C.1 Approximating the $\mathcal{R}\mathrm{TV}^2$ for sigmoidal smooth decision trees

For the sake of clarity, in Section 5, we analyse the function

$$f_{\mathsf{DT},\gamma}(\boldsymbol{x}) \;=\; \prod_{i=1}^{D} \sigma_\gamma\big(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i\big), \qquad \boldsymbol{x} \in \mathbb{R}^d,\ \gamma > 0, \tag{12}$$

where

$$\sigma_\gamma(z) \;=\; \frac{1}{1 + e^{-z/\gamma}}$$

is a *scaled logistic*, and $\boldsymbol{w}_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ are the split normals and thresholds at depth i.

Now, we show the proof of the result on the Fourier transform of Sigmoidal smooth decision trees. First, we restate the result on the Fourier transform of a shifted, scaled sigmoid, then provide the proof below.

Lemma 11 *For any $\gamma > 0$ and $b \in \mathbb{R}$,*

$$\widehat{\sigma_\gamma(\cdot + b)}(\omega) \;=\; e^{-\mathrm{i}\,\omega b}\Big[\frac{\pi}{2}\,\delta(\omega) + \frac{\mathrm{i}\gamma\pi}{\sinh(\pi\gamma\omega)}\Big], \qquad \omega \in \mathbb{R}.$$

Proof

First, note that, by translation invariance,

$$\widehat{\sigma_\gamma(\cdot + b)}(\omega) = e^{-\mathrm{i}\omega b}\,\widehat{\sigma_\gamma}(\omega),$$

so it suffices to compute $\widehat{\sigma_\gamma}$.

We can write the sigmoid as

$$\sigma_\gamma(z) = \tfrac{1}{2} + \tfrac{1}{2}\tanh\Big(\tfrac{z}{2\gamma}\Big).$$

For the constant part $\widehat{\frac{1}{2}} = \dfrac{\pi}{2}\delta(\omega)$.

For the hyperbolic–tangent part use the table entry

$$\int_{-\infty}^{\infty} \tanh u\, e^{-\mathrm{i}\Omega u}\, du = \frac{\pi\mathrm{i}}{\sinh(\pi\Omega/2)} \qquad \text{(Bateman Vol 1, §4.9, (9) (Project et al., 1954))},$$

together with the scaling rule $\widehat{f(ax)}(\omega) = \dfrac{1}{|a|}\widehat{f}\big(\frac{\omega}{a}\big)$ for $a \neq 0$. Choosing $a = 2\gamma$ gives

$$\widehat{\tanh(\cdot/2\gamma)}(\omega) = 2\gamma\,\frac{\pi\mathrm{i}}{\sinh(\pi\gamma\omega)}.$$

Combining the two terms and re-inserting the phase factor gives- for all $\omega \in \mathbb{R}$

$$\widehat{\sigma_\gamma(\cdot + b)}(\omega) = e^{-\mathrm{i}\omega b}\Big[\frac{\pi}{2}\,\delta(\omega) + \frac{\mathrm{i}\gamma\pi}{\sinh(\pi\gamma\omega)}\Big],$$

as claimed. ■

In the following subsection, we consider the geometry of the splits direction $\{\boldsymbol{w}_i\}_{i=1}^{D} \subset \mathbb{R}^d$.

C.2 Geometry of the split directions

First, consider rewriting all the split normals into a matrix

$$W \;:=\; \begin{bmatrix} \boldsymbol{w}_1^\top \\ \vdots \\ \boldsymbol{w}_D^\top \end{bmatrix} \in \mathbb{R}^{D\times d},$$

Denote by

$$r \;:=\; \operatorname{rank} W \qquad (0 \le r \le \min\{D, d\})$$

the dimension of their span

$$\mathcal{S} \;:=\; \operatorname{span}\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D\} \subset \mathbb{R}^d.$$

C.3 Rotating into $\mathcal{S} \oplus \mathcal{S}^\perp$

Now, we will rewrite the product in the expression of Sigmoidal smooth decision tree in Eq. (3) corresponding to the active directions in the span of the split normals $\mathcal{S}$.

First, note that $\mathcal{S}$ is r–dimensional, so we can pick an orthogonal matrix $\boldsymbol{R} = [\boldsymbol{R}_\parallel \;\; \boldsymbol{R}_\perp]$ such that

$$\operatorname{im} \boldsymbol{R}_\parallel = \mathcal{S}, \quad \operatorname{im} \boldsymbol{R}_\perp = \mathcal{S}^\perp.$$

Notation. We write every point $\boldsymbol{x} \in \mathbb{R}^d$ and every frequency $\boldsymbol{\xi}$ in these coordinates:

$$\boldsymbol{x} = \boldsymbol{R}\begin{bmatrix} \boldsymbol{x}_\parallel \\ \boldsymbol{x}_\perp \end{bmatrix}, \quad \boldsymbol{x}_\parallel \in \mathbb{R}^r,\ \boldsymbol{x}_\perp \in \mathbb{R}^{d-r}; \qquad \boldsymbol{\xi} = \boldsymbol{R}\begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{\zeta} \end{bmatrix}, \quad \boldsymbol{\eta} \in \mathbb{R}^r,\ \boldsymbol{\zeta} \in \mathbb{R}^{d-r}.$$

Now, because each $\boldsymbol{w}_i$ lies *inside* $\mathcal{S}$ we have $\boldsymbol{w}_i^\top \boldsymbol{R}_\perp = 0$; hence

$$\boldsymbol{w}_i^\top \boldsymbol{x} + b_i = \boldsymbol{w}_i^\top \boldsymbol{R}\begin{bmatrix} \boldsymbol{x}_\parallel \\ \boldsymbol{x}_\perp \end{bmatrix} + b_i = \underbrace{(\boldsymbol{w}_i^\top \boldsymbol{R}_\parallel)}_{=:\boldsymbol{a}_i^\top}\boldsymbol{x}_\parallel + b_i, \quad \boldsymbol{a}_i \in \mathbb{R}^r. \tag{13}$$

Thus $f_{\mathsf{DT},\gamma}$ depends *only* on the $\boldsymbol{x}_\parallel$–coordinates:

$$f_{\mathsf{DT},\gamma}\big(\boldsymbol{R}[\boldsymbol{x}_\parallel, \boldsymbol{x}_\perp]^\top\big) = \prod_{i=1}^{D} \sigma_\gamma(\boldsymbol{a}_i^\top \boldsymbol{x}_\parallel + b_i).$$

C.4 Splitting the Fourier integral

By definitions, we know that the Fourier transform of $f_{\mathsf{DT},\gamma}$ is given by

$$\widehat{f_{\mathsf{DT},\gamma}}(\xi) = (2\pi)^{-d/2} \int f_{\mathsf{DT},\gamma}(\boldsymbol{x}) e^{-\mathrm{i}\boldsymbol{\xi}^\top \boldsymbol{x}}\, d\boldsymbol{x} = (2\pi)^{-d/2} \int \Big[\prod_{i=1}^{D} \sigma_\gamma(\boldsymbol{a}_i^\top \boldsymbol{x}_\parallel + b_i)\Big] \cdot e^{-\mathrm{i}\boldsymbol{\xi}^\top \boldsymbol{x}}\, d\boldsymbol{x} \tag{14}$$

Now, insert the rotated coordinates into the definition

$$\widehat{f_{\mathsf{DT},\gamma}}(\xi) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \Big[\prod_{i=1}^{D} \sigma_\gamma(a_i^\top t + b_i)\Big]\, e^{-\mathrm{i}\,(\eta^\top t + \zeta^\top z)}\, dt\, dz$$
$$= (2\pi)^{-d/2} \underbrace{\Big[\int_{\mathbb{R}^{d-r}} e^{-\mathrm{i}\,\zeta^\top z}\, dz\Big]}_{=\,(2\pi)^{(d-r)/2}\delta(\zeta)} \int_{\mathbb{R}^r} G(t)\, e^{-\mathrm{i}\,\eta^\top t}\, dt, \tag{15}$$

where we define $G(t) := \prod_{i=1}^{D} \sigma_\gamma(a_i^\top t + b_i)$.

The outer integral in Eq. (15) produced a *Dirac delta* $\delta(\zeta) = \delta(P_{\mathcal{S}^\perp}\xi)$ that kills every frequency component *outside* $\mathcal{S}$; we are left with an r-dimensional Fourier transform of G inside $\mathcal{S}$.

Then, Eq. (15) becomes

$$\widehat{f_{\mathsf{DT},\gamma}}(\xi) = (2\pi)^{-r/2}\, \delta\big(P_{\mathcal{S}^\perp}\xi\big) \underbrace{\int_{\mathbb{R}^r} G(t)\, e^{-\mathrm{i}\,\eta^\top t}\, dt}_{=:H(\eta)}. \tag{16}$$

C.5 Expressing $H(\eta)$ as a D-fold convolution

Now, consider the linear projection of t wrt the matrix A: $t \mapsto s := At$, where A is formed as the *row-stacked* matrix

$$A := \begin{bmatrix} a_1^\top \\ \vdots \\ a_D^\top \end{bmatrix} \in \mathbb{R}^{D\times r}, \qquad a_i^\top = \boldsymbol{w}_i^\top R_\parallel\ (1 \times r).$$

Because $\operatorname{rank} A = r$ (same as W), the linear map $t \mapsto s := At$ is injective, sending $t \in \mathbb{R}^r$ to a vector $s = (s_1, \ldots, s_D)^\top \in \mathbb{R}^D$ whose i-th entry is $s_i = a_i^\top t$. In these s–coordinates the kernel factorises:

$$G(t) = \prod_{i=1}^{D} \sigma_\gamma\big(s_i + b_i\big).$$

Solve for t by left-multiplying with the Moore–Penrose inverse $A^+ = (A^\top A)^{-1}A^\top \in \mathbb{R}^{r\times D}$:

$$t = A^+ s, \qquad dt = \big|\det(A^\top A)\big|^{-1/2}\, ds.$$

Note that $A^\top A$ is $r \times r$, so the determinant is well defined. With $t = A^+ s$ we have

$$\eta^\top t = \eta^\top (A^\top A)^{-1} A^\top s = \underbrace{\big(A(A^\top A)^{-1}\eta\big)^\top}_{=:u^\top} s, \quad u := A(A^\top A)^{-1}\eta\ \in \mathbb{R}^D.$$

Now, we show one-dimensional convolution in each coordinate. Insert these expressions into $H(\eta)$:

$$H(\eta) = \big|\det(A^\top A)\big|^{-1/2} \int_{\mathbb{R}^D} \Big[\prod_{i=1}^{D} \sigma_\gamma(s_i + b_i)\Big]\, e^{-\mathrm{i}\,u^\top s}\, ds \tag{17}$$

$$\overset{\text{F.T.}}{=} \big|\det(A^\top A)\big|^{-1/2} \Big(K_{\gamma,1} * K_{\gamma,2} * \cdots * K_{\gamma,D}\Big)(u), \tag{18}$$

where $K_{\gamma,i} = \widehat{\sigma_\gamma(\cdot + b_i)}$ (from Lemma 11). Each convolution is in the $\mathbb{R}$–variable corresponding to the i-th coordinate.

The change of variables $t \mapsto s$ contributed the Jacobian $|\det(A^\top A)|^{-1/2}$. Because $A = WR_\parallel$ and $R_\parallel$ is orthogonal on $\mathcal{S}$, one has $A^\top A = R_\parallel^\top (W^\top W) R_\parallel = W^\top W|_{\mathcal{S}}$, so $|\det(A^\top A)| = |\det(WW^\top)|$. Define the *left* pseudo-inverse

$$W^+ := (WW^\top)^{-1} W \in \mathbb{R}^{D\times d}.$$

Combining (16) with (18) now yields the following arbitrary split formula.

Theorem 12 (Fourier transform, arbitrary split directions) *Let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_D \in \mathbb{R}^d$, $b_1, \dots, b_D \in \mathbb{R}$ and $\gamma > 0$. Put $W = \begin{bmatrix}\boldsymbol{w}_1 & \cdots & \boldsymbol{w}_D\end{bmatrix}^\top \in \mathbb{R}^{D\times d}$, $r = \operatorname{rank} W$, $\mathcal{S} = \operatorname{span}\{\boldsymbol{w}_i\}$. With the unitary convention*

$$\widehat{g}(\xi) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} g(x)\, e^{-\mathrm{i}\xi^\top x}\, dx,$$

the the Sigmoid smooth decision tree with depth $D > 0$

$$f_{DT,\gamma}(x) = \prod_{i=1}^{D} \sigma_\gamma(\boldsymbol{w}_i^\top x + b_i)$$

has Fourier transform

$$\widehat{f_{DT,\gamma}}(\xi) = (2\pi)^{-r/2}\, |\det(WW^\top)|^{-1/2}\, \delta\big(P_{\mathcal{S}^\perp}\xi\big)\, \Big(K_{\gamma,1} * \cdots * K_{\gamma,D}\Big)\big(W^+\xi\big),$$

where $K_{\gamma,i}(u) = e^{-\mathrm{i}b_i u}\big[\frac{\pi}{2}\delta(u) + \frac{\mathrm{i}\gamma\pi}{\sinh(\pi\gamma u)}\big]$ is the 1-D kernel from Lemma 11.

Remark 13 *In the theorem above, we assumed that the split normals $\{\boldsymbol{w}_i\}$ are distinct. In the case when the directions are not distinct, one can consider a maximally linearly independent set of split normals to obtain similar results as above. Thus, the convolution of the 1-D kernels involves multiplicity of the split normals in the depth product.*

Suppose now that the $\boldsymbol{w}_i$ are *orthonormal and distinct*. Then $r = D \le d$ and $W^+ = W^\top$, $WW^\top = I_D$, $A = I_D$. Because each $K_{\gamma,i}$ acts on an *independent coordinate* (the i-th standard basis vector), convolutions reduce to ordinary point-wise products:

$$K_{\gamma,1} * \cdots * K_{\gamma,D} = K_{\gamma,1} \cdots K_{\gamma,D}.$$

This yields the following product-form for orthonormal splits in the Fourier transform.

Corollary 14 (Orthonormal $\{\boldsymbol{w}_i\}$, $D \le d$)

$$\widehat{f_{DT,\gamma}}(\xi) = (2\pi)^{-D/2}\, \delta\big(P_{\mathcal{S}^\perp}\xi\big) \prod_{i=1}^{D} e^{-\mathrm{i}\, b_i\eta_i} \Big[\frac{\pi}{2}\,\delta(\eta_i) + \frac{\mathrm{i}\,\gamma\pi}{\sinh(\pi\gamma\eta_i)}\Big], \quad \eta_i = \boldsymbol{w}_i^\top \xi. \tag{19}$$

With this we provide the proof of the main theorem.

C.6 Proof of Theorem 6

Notation. Denote by

$$\lambda_i(\beta) = \boldsymbol{w}_i^\top \beta \quad (i = 1, \dots, D),$$

and write $d\sigma(\beta)$ for the surface measure on the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$.

Note that, the tempered distribution $f_{\mathsf{DT},\gamma}$ has Fourier transform

$$\widehat{f_{\mathsf{DT},\gamma}}(\xi) = (2\pi)^{-\frac{D}{2}}\, \delta\big(P_{\mathcal{S}^\perp}\xi\big) \prod_{i=1}^{D} e^{-\mathrm{i}\, b_i \boldsymbol{w}_i^\top \xi} \Big[\frac{\pi}{2}\, \delta\big(\boldsymbol{w}_i^\top \xi\big) + \frac{\mathrm{i}\,\gamma\pi}{\sinh\big(\pi\gamma \boldsymbol{w}_i^\top \xi\big)}\Big],$$

so $\widehat{f_{\mathsf{DT},\gamma}}$ is supported on $\mathcal{S}$.

Now, evaluating the Fourier transform on the $\omega\boldsymbol{\beta}$ we get

$$\widehat{f_{\mathsf{DT},\gamma}}(\omega\boldsymbol{\beta}) = (2\pi)^{-\frac{D}{2}}\, \delta\big(P_{\mathcal{S}^\perp}\omega\boldsymbol{\beta}\big) \prod_{i=1}^{D} e^{-\mathrm{i}\, b_i \omega \boldsymbol{w}_i^\top \boldsymbol{\beta}} \Big[\frac{\pi}{2}\, \delta\big(\omega \boldsymbol{w}_i^\top \boldsymbol{\beta}\big) + \frac{\mathrm{i}\,\gamma\pi}{\sinh\big(\pi\gamma\omega \boldsymbol{w}_i^\top \boldsymbol{\beta}\big)}\Big]$$

Now, we can write the 1D inverse Fourier transform to solve for the Radon transform of $f_{\mathsf{DT},\gamma}$

$$\mathcal{R}\left\{f_{\mathsf{DT},\gamma}\right\}(\boldsymbol{\beta}, t) = (2\pi)^{-1/2} \int_{\mathbb{R}} e^{\mathrm{i}\omega t} \widehat{f_{\mathsf{DT},\gamma}}(\omega\boldsymbol{\beta})\, d\omega$$

Because of the factor $\delta\big(P_{\mathcal{S}^\perp}\xi\big)$ in Eq. (19), the integrand is non–zero *only if* $\beta_\perp = 0$, i.e. $\beta \in \mathcal{S}$. Hence

$$\mathcal{R}\left\{f_{\mathsf{DT},\gamma}\right\}(\boldsymbol{\beta}, t) = 0 \quad \text{unless} \quad \beta \in \mathcal{S} \cap \mathbb{S}^{d-1}.$$

For the rest of the discussion we assume that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_D, 0, \dots, 0)$ with (almost surely) $\boldsymbol{\beta}_i \neq 0$ for all $i \in [D]$.

Because $\widehat{f_{\mathsf{DT},\gamma}}(\omega\boldsymbol{\beta})$ has at most polynomial growth in ω, we may differentiate under the integral:

$$\partial_t^{d+1} \mathcal{R}\left\{f_{\mathsf{DT},\gamma}\right\}(\boldsymbol{\beta}, t) = (2\pi)^{-1/2} \int_{\mathbb{R}} (\mathrm{i}\omega)^{d+1}\, e^{\mathrm{i}\omega t}\, \widehat{f_{\mathsf{DT},\gamma}}(\omega\boldsymbol{\beta})\, d\omega. \tag{20}$$

Now, integrating absolute value of LHS in Eq. (20),

$$\begin{aligned} ||\partial_t^{d+1} \mathcal{R}\left\{f_{\mathsf{DT},\gamma}\right\}(\boldsymbol{\beta}, \cdot)||_{L_t^1} &= (2\pi)^{-1/2} \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} (\mathrm{i}\omega)^{d+1}\, e^{\mathrm{i}\omega t}\, \widehat{f_{\mathsf{DT},\gamma}}(\omega\boldsymbol{\beta})\, d\omega \Big|\, dt \\ &= (2\pi)^{-1/2} \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} (\omega)^{d+1}\, e^{\mathrm{i}\omega t}\, \widehat{f_{\mathsf{DT},\gamma}}(\omega\boldsymbol{\beta})\, d\omega \Big|\, dt \end{aligned} \tag{21}$$

Now, note that expanding the full form of $\widehat{f_{\mathsf{DT},\gamma}}$ in Eq. (21), we can eliminate the $\frac{\pi}{2}\delta(\omega \boldsymbol{w}_i^\top \boldsymbol{\beta})$ terms due to the application of the Sifting property of Dirac delta on $|\omega|^{d+1}\delta(\omega)$. Hence, we can simplify the Eq. (21) as follows:

$$||\partial_t^{d+1} \mathcal{R}\left\{f_{\mathsf{DT},\gamma}\right\}(\boldsymbol{\beta}, \cdot)||_{L_t^1} = (2\pi)^{-(D+1)/2} (\gamma\pi)^D \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} (\omega)^{d+1}\, e^{\mathrm{i}\omega\left(t - \sum_{i=1}^{D} b_i \boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \prod_{i=1}^{D} \frac{1}{\sinh\big(\pi\gamma\omega \boldsymbol{w}_i^\top \boldsymbol{\beta}\big)}\, d\omega \Big|\, dt \tag{22}$$

Now, integrating Eq. (21) with respect to $\boldsymbol{\beta}$ yields the $\mathcal{R}\mathrm{TV}^2$ of $f_{\mathsf{DT},\gamma}$,

$$\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT},\gamma}) = c_D \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} (\omega)^{d+1} \, e^{\mathrm{i}\omega\left(t - \sum_{i=1}^{D} b_i \boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \prod_{i=1}^{D} \frac{1}{\sinh\left(\pi\gamma\omega \boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \, d\omega \Big| \, dt \, d\boldsymbol{\beta}$$

where $c_D := (2\pi)^{-(D+1)/2}(\gamma\pi)^D$. Since $\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT},\gamma})$ is non-zero only for $\boldsymbol{\beta} \in \S \cap \mathbb{S}^{d-1}$, we can further simplify the equation above as

$$\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT},\gamma}) = c_D \int_{\mathbb{S}^{D-1}} \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} (\omega)^{d+1} \, e^{\mathrm{i}\omega\left(t - \sum_{i=1}^{D} b_i \boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \prod_{i=1}^{D} \frac{1}{\sinh\left(\pi\gamma\omega \boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \, d\omega \Big| \, dt \, d\boldsymbol{\beta}$$

Finally, we will show that the RHS obtained in the form of $\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT},\gamma})$ above is unbounded.

First, we note that for any $\varepsilon > 0$, the following integral is bounded:

$$I_\varepsilon[\varepsilon, \infty] := \int_{\varepsilon}^{\infty} \omega^{d+1} \, e^{\mathrm{i}\omega\left(t - \sum_{i=1}^{D} b_i \, \boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \prod_{i=1}^{D} \frac{1}{\sinh\left(\pi\gamma\omega \, \boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \, d\omega \tag{23}$$

Note that for all $i = 1, 2, \ldots, D$, almost surely $\boldsymbol{w}_i^\top \boldsymbol{\beta} \neq 0$. Now, by definition for large $z > 0$, $\sinh z$ decays like $\frac{1}{e^z}$. Thus, for bound t values, $I_\varepsilon[\varepsilon, \infty]$ is bounded by a constant. Similarly, we can bound $I_\varepsilon[\infty, -\varepsilon]$.

Now, we will show a construction of a cone in $\mathbb{S}^{D-1}$ and an interval of t such that $\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT},\gamma})$ is lower bounded by a quantity that turns out to be unbounded.

For $\varepsilon > 0$ small enough, without loss of generality pick $i = 1$ and set

$$E_\varepsilon := \left\{ \boldsymbol{\beta} \in \mathbb{S}^{D-1} : |\boldsymbol{w}_1^\top \boldsymbol{\beta}| < \varepsilon, \quad |\boldsymbol{w}_j^\top \boldsymbol{\beta}| \geq c_0 > 0 \ (j \geq 2) \right\},$$

where $c_0 := \min_{j \geq 2, \, \boldsymbol{\beta} \in \mathbb{S}^{D-1}} |\boldsymbol{w}_j^\top \boldsymbol{\beta}| > 0$ after a possible renormalisation. Note that $\mu(E_\varepsilon) = \Omega(\varepsilon)$.

For $|\omega| \leq \varepsilon$ and $\boldsymbol{\beta} \in E_\varepsilon$ we have

$$\left|\sinh(\pi\gamma\omega \, \boldsymbol{w}_1^\top \boldsymbol{\beta})\right| \leq |\pi\gamma\omega \, \boldsymbol{w}_1^\top \boldsymbol{\beta}|, \qquad \left|\sinh(\pi\gamma\omega \, \boldsymbol{w}_j^\top \boldsymbol{\beta})\right| \leq |\pi\gamma\omega| \quad (j \geq 2).$$

Thus, for the choice of $\boldsymbol{\beta}$ and ω

$$\prod_{i=1}^{D} \frac{1}{\left|\sinh(\pi\gamma\omega \, \boldsymbol{w}_i^\top \boldsymbol{\beta})\right|} \geq K \, \frac{1}{|\beta_1|} \, \frac{1}{|\omega|^D} \tag{24}$$

for some constant $K > 0$. Now, consider the integral $I_\varepsilon[-\varepsilon, \varepsilon]$,

$$\begin{aligned}
I_\varepsilon[-\varepsilon, \varepsilon] &= \int_{-\varepsilon}^{\varepsilon} \omega^{d+1} \, e^{\mathrm{i}\omega\delta_{\boldsymbol{\beta}}(t)} \prod_{i=1}^{D} \frac{1}{\sinh\left(\pi\gamma\omega \, \boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \, d\omega \\
&= \int_{0}^{\varepsilon} \omega^{d+1} \left(e^{\mathrm{i}\omega\delta_{\boldsymbol{\beta}}(t)} + (-1)^{d+1-D} e^{-\mathrm{i}\omega\delta_{\boldsymbol{\beta}}(t)} \right) \prod_{i=1}^{D} \frac{1}{\sinh\left(\pi\gamma\omega \, \boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \, d\omega \\
&= c_{\pi,\gamma} \int_{0}^{\varepsilon} \omega^{d+1-D} \underbrace{\left(e^{\mathrm{i}\omega\delta_{\boldsymbol{\beta}}(t)} + (-1)^{d+1-D} e^{-\mathrm{i}\omega\delta_{\boldsymbol{\beta}}(t)} \right)}_{(\bigstar)} \prod_{i=1}^{D} \frac{1}{\left(\boldsymbol{w}_i^\top \boldsymbol{\beta}\right)} \, d\omega
\end{aligned}$$

where $\delta_{\boldsymbol{\beta}}(t) := t - \sum_i b_i \, \boldsymbol{w}_i^\top \boldsymbol{\beta}$ and $c_{\pi,\gamma}$ is a constant that depends on π and γ. Now, irrespective of the choice of d and D, we have

$$(\bigstar) = \begin{cases} 2\cos\left(\omega \delta_{\boldsymbol{\beta}}(t)\right) & \text{if } d+1-D \equiv 0 \mod 2 \\ 2\mathrm{i}\sin\left(\omega \delta_{\boldsymbol{\beta}}(t)\right) & \text{if } d+1-D \equiv 1 \mod 2 \end{cases} \tag{25}$$

Choose the slice $T_\varepsilon(\boldsymbol{\beta}) := \{\, t : |\delta_{\boldsymbol{\beta}}(t)| < \varepsilon^2 \}$. For $t \in T_\varepsilon(\boldsymbol{\beta})$ and $\omega \le \varepsilon$ we have $|\omega \delta_{\boldsymbol{\beta}}(t)| \le \varepsilon^3 \ll 1$, so $|\cos(\omega \delta_{\boldsymbol{\beta}}(t))| \ge \frac{1}{2}$ and $|\sin(\omega \delta_{\boldsymbol{\beta}}(t))| \ge \frac{2}{\pi}\omega |\delta_{\boldsymbol{\beta}}(t)|$ (for the later we need $\omega |\delta_{\boldsymbol{\beta}}(t)| \le \varepsilon^3$)

Now, we show the unboundedness of $I_\varepsilon[-\varepsilon, \varepsilon]$ for the case when $d + 1 - D \equiv 0 \mod 2$; whereas the other case follows similarly.

$$|g_{\boldsymbol{\beta}}(t)| := \Big| \int_0^\varepsilon \omega^{d+1} e^{\mathrm{i}\omega(t - \sum b_i \, \boldsymbol{w}_i^\top \boldsymbol{\beta})} \prod_{i=1}^{D} \frac{1}{\sinh(\pi\gamma\omega \, \boldsymbol{w}_i^\top \boldsymbol{\beta})} \, d\omega \Big| \;\ge\; \frac{K}{|\beta_1|} \int_0^\varepsilon \omega^{d+1-D} \left|\cos\big(\omega \delta_{\boldsymbol{\beta}}(t)\big)\right| d\omega,$$

where we have used the lower bound from Eq. (24) in the RHS. But noting the boundedness of $\cos\big(\omega \delta_{\boldsymbol{\beta}}(t)\big)$, we can write

$$|g_{\boldsymbol{\beta}}(t)| \;\ge\; \frac{K'}{|\beta_1|} \int_0^\varepsilon \omega^{d+1-D} \, d\omega \;=\; \frac{K'' \, \varepsilon^{d+2-D}}{|\beta_1|}.$$

The set $T_\varepsilon(\boldsymbol{\beta})$ has length at most $2\varepsilon^2$, giving

$$\int_{\mathbb{R}} |g_{\boldsymbol{\beta}}(t)| \, dt \;\ge\; \frac{C_1 \, \varepsilon^{d+4-D}}{|\beta_1|},$$

for some constant $C_1 > 0$. Integrating over E_ε,

$$\int_{E_\varepsilon} \int_{\mathbb{R}} |g_{\boldsymbol{\beta}}(t)| \, dt \, d\boldsymbol{\beta} \ge \int_{E_\varepsilon} \frac{C_1 \, \varepsilon^{d+4-D}}{|\beta_1|} d\boldsymbol{\beta} \to \infty.$$

Hence the entire triple integral involved in $\mathcal{R}\mathrm{TV}^2(f_{\mathsf{DT},\gamma})$.

Appendix D. Gaussian Smoothing

In this appendix, we provide the proof of the main theorem on the approximation of Gaussian smoothed decision trees as stated in Theorem 7.

D.1 Upper bounding the $\mathcal{R}\mathrm{TV}^2$ for the Gaussain-smoothed decision trees

Let $f_{\mathsf{DT}}(\cdot) = 1\{\cdot \in A\}$ be the indicator function of a decision-tree region $A \subset \mathbb{R}^d$, and define the smoothed function via convolution:

$$f_\sigma(\boldsymbol{x}) := (f_{\mathsf{DT}} * G_\sigma)(\boldsymbol{x}) = \int_{\mathbb{R}^d} 1\{\boldsymbol{y} \in A\}\; G_\sigma(\boldsymbol{x}-\boldsymbol{y})\, d\boldsymbol{y},$$

where $G_\sigma(\boldsymbol{z}) := \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\boldsymbol{z}\|^2}{2\sigma^2}\right)$ is the Gaussian kernel.

By the Fourier-slice formula,

$$R\{f_\sigma\}(\boldsymbol{\beta}, t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\mathrm{i}\omega t} \widehat{f_\sigma}(\omega\boldsymbol{\beta})\, d\omega,$$

and since $\widehat{f_\sigma}(\boldsymbol{\xi}) = \widehat{f_{\mathsf{DT}}}(\boldsymbol{\xi})\, e^{-\frac{\sigma^2\|\boldsymbol{\xi}\|^2}{2}}$ (using Convolution theorem of Fourier transform), we compute:

$$\partial_t^{d+1} R\{f_\sigma\}(\boldsymbol{\beta}, t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} (\mathrm{i}\omega)^{d+1} e^{\mathrm{i}\omega t} \widehat{f_{\mathsf{DT}}}(\omega\boldsymbol{\beta})\, e^{-\frac{\sigma^2\omega^2}{2}} d\omega. \tag{26}$$

As shown in Eq. (9), we have:

$$\widehat{f_{\mathsf{DT}}}(\omega\boldsymbol{\beta}) = \frac{1}{(2\pi)^{d/2}} \int_A e^{-\mathrm{i}\omega\boldsymbol{\beta}^\top \boldsymbol{x}} d\boldsymbol{x}.$$

Plugging this equation in Eq. (26), we get:

$$\partial_t^{d+1} R\{f_\sigma\}(\boldsymbol{\beta}, t) = \frac{1}{(2\pi)^{(d+1)/2}} \int_A \underbrace{\left[\int_{\mathbb{R}} (\mathrm{i}\omega)^{d+1} e^{\mathrm{i}\omega(t-\boldsymbol{\beta}^\top\boldsymbol{x})} e^{-\frac{\sigma^2\omega^2}{2}} d\omega\right]}_{I_{d+1}} d\boldsymbol{x}. \tag{27}$$

In the following, we would rewrite I_{d+1} in terms of probabilist's Hermite polynomials.

Set $m = d+1$ and $s := t - \boldsymbol{\beta}^\top\boldsymbol{x}$. Now, using Rodrigues' formula for Hermite polynomials: $\mathrm{He}_n(u) = (-1)^n e^{\frac{u^2}{2}} \frac{d^n}{du^n} e^{-\frac{u^2}{2}}$, and integration by parts, we have

$$I_m = \int_{\mathbb{R}} (\mathrm{i}\omega)^m\, e^{\mathrm{i}\omega s}\, e^{-\frac{\sigma^2\omega^2}{2}}\, d\omega = (-\mathrm{i})^m\, \frac{\sqrt{2\pi}}{\sigma^{m+1}}\, e^{-\frac{s^2}{2\sigma^2}}\, \mathrm{He}_m\Big(\frac{s}{\sigma}\Big). \tag{28}$$

With this, we can write:

$$\left|\partial_t^{d+1} R\{f_\sigma\}(\boldsymbol{\beta}, t)\right| = \frac{1}{(2\pi)^{(d+1)/2}} \left|\int_A (-\mathrm{i})^{d+1}\, \frac{\sqrt{2\pi}}{\sigma^{d+2}}\, e^{-\frac{(t-\boldsymbol{\beta}^\top\boldsymbol{x})^2}{2\sigma^2}}\, \mathrm{He}_{d+1}\Big(\frac{t-\boldsymbol{\beta}^\top\boldsymbol{x}}{\sigma}\Big) d\boldsymbol{x}\right| \tag{29}$$

Hence, we can write

$$\left|\partial_t^{d+1} R\{f_\sigma\}(\boldsymbol{\beta}, t)\right| \le \frac{\sigma^{-(d+2)}}{(2\pi)^{d/2}} \int_A \left|\mathrm{He}_{d+1}\Big(\tfrac{t-\boldsymbol{\beta}^\top\boldsymbol{x}}{\sigma}\Big)\right| \exp\Big(-\tfrac{(t-\boldsymbol{\beta}^\top\boldsymbol{x})^2}{2\sigma^2}\Big)\, d\boldsymbol{x}. \tag{30}$$

In odd dimensions, $\mathcal{R}\text{TV}^2$ of f_σ is the L^1-norm of the absolute value of $(d+1)-$th derivative of $R\{f_\sigma\}(\boldsymbol{\beta}, t)$ with respect to t. Thus, we get

$$\begin{aligned}
||f_\sigma||_{\mathcal{R}} &= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left|\partial_t^{d+1} R\{f_\sigma\}(\boldsymbol{\beta}, t)\right| dt\, d\boldsymbol{\beta} \\
&\leq \frac{\sigma^{-(d+2)}}{(2\pi)^{d/2}} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \int_A \left|\text{He}_{d+1}\Big(\tfrac{t-\boldsymbol{\beta}^\top \boldsymbol{x}}{\sigma}\Big)\right| \exp\Big(-\tfrac{(t-\boldsymbol{\beta}^\top \boldsymbol{x})^2}{2\sigma^2}\Big)\, d\boldsymbol{x}\, dt\, d\boldsymbol{\beta} \\
&= \frac{\sigma^{-(d+2)}}{(2\pi)^{d/2}} \int_{\mathbb{S}^{d-1}} \int_A \int_{\mathbb{R}} \left|\text{He}_{d+1}\Big(\tfrac{t-\boldsymbol{\beta}^\top \boldsymbol{x}}{\sigma}\Big)\right| \exp\Big(-\tfrac{(t-\boldsymbol{\beta}^\top \boldsymbol{x})^2}{2\sigma^2}\Big)\, dt\, d\boldsymbol{x}\, d\boldsymbol{\beta}
\end{aligned}$$

Substituting $u = \frac{t-\boldsymbol{\beta}^\top \boldsymbol{x}}{\sigma}$, $dt = \sigma du$, and noting that this inner integral is independent of x, we obtain:

$$||f_\sigma||_{\mathcal{R}} \leq \frac{\sigma^{-(d+1)}}{(2\pi)^{d/2}} \int_{\mathbb{S}^{d-1}} \int_A C_{\text{He}}(d+1) d\boldsymbol{x}\, d\boldsymbol{\beta} \tag{31}$$

where we define

$$C_{\text{He}}(m) \;:=\; \int_{\mathbb{R}} e^{-u^2/2} \left|\text{He}_m(u)\right| du.$$

Simplifying the integrals with respect to t and $\boldsymbol{\beta}$, we can rewrite Eq. (31) as

$$||f_\sigma||_{\mathcal{R}} \leq \frac{\sigma^{-(d+1)}}{(2\pi)^{d/2}} C_{\text{He}}(d+1) \cdot Vol(A) \cdot |\mathbb{S}^{d-1}| \tag{32}$$

Now, we will bound $C_{\text{He}}(d+1)$. Note that

$$C_{\text{He}}(m) \;=\; \int_{\mathbb{R}} e^{-u^2/2} \left|\text{He}_m(u)\right| du \leq \sqrt{\left(\int_{\mathbb{R}} e^{-u^2/2}\, du\right)\left(\int_{\mathbb{R}} e^{-u^2/2} \,\text{He}_m^2(u)\, du\right)} \quad (\star)$$

where we have used Cauchy-Schwarz in the last inequality.

Note that the first term in $(\star)$ has a concrete form due to integral of a Gaussian density, and hence

$$\int_{\mathbb{R}} e^{-u^2/2}\, du = \sqrt{2\pi}$$

In the second term of the $(\star)$, we use a standard identity on inner product of probabilit's Hermite polynomials,

$$\int_{\mathbb{R}} e^{-u^2/2} \,\text{He}_m^2(u)\, du = \sqrt{2\pi} m!$$

Using the approximations above, we can bound

$$C_{\text{He}}(d+1) \leq \sqrt{2\pi}\sqrt{(d+1)!}$$

Now, using Sterling's approximation, we can further simplify the rhs as

$$C_{\text{He}}(d{+}1) \;\leq\; c\,(d{+}1)^{\frac{d+1}{2}}\; 2^{d/2} \qquad (\text{some universal } c \approx 1.2).$$

for $d \geq 3$.

Using this bound to simplify Eq. (32), we get

$$\begin{aligned} ||f_\sigma||_{\mathcal{R}} &\leq \frac{\sigma^{-(d+1)}}{(2\pi)^{d/2}} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot c\,(d{+}1)^{\frac{d+1}{2}}\ 2^{d/2} \cdot Vol(A) \\ &= \frac{\sigma^{-(d+1)}}{\Gamma(d/2)} \cdot c\,(d{+}1)^{\frac{d+1}{2}} \cdot Vol(A) \end{aligned}$$

Now, using the Stirling's approximation on $\Gamma(d/2) \approx \sqrt{2\pi}\left(\frac{d}{2e}\right)^{d/2}\sqrt{\frac{2}{d}}$: we have

$$\|f_\sigma\|_R \ \leq\ C\,d^{1/2}\left(\frac{\sqrt{2}\,e}{\sigma}\right)^d Vol(A) \quad (\sigma > 0),$$

where $C \leq 2.2$.

Appendix E. Approximation Post-thresholding

In this section we prove Lemma 8 and Theorem 9 from Section 6.

Proof [Proof of Lemma 8]

Exact thresholding. If $x \in B$, then for each coordinate $t \in \{u_j - x_j,\ x_j - \ell_j\} \geq 0$ and $\vartheta_{\lambda,\varepsilon}(t) = 1$; hence $S_B(x) = 1$. If $x \notin B$, then at least one factor is < 1, so $S_B(x) < 1$. Thus $\{S_B \geq 1\} = B$.

Distributional closeness. Write the coordinate overhangs $\delta_j(x) := (\ell_j - x_j)_+ + (x_j - u_j)_+$, so $Z = d_1(x, B) = \sum_{j=1}^d \delta_j(x)$ and put $Z_2 := \operatorname{dist}(x, \partial B)$ (Euclidean). For $t \leq 0$ the barrier satisfies $\vartheta_{\lambda,\varepsilon}(t) \leq 1$ and, if $t \leq -\varepsilon$, equals $e^{\lambda t}$. Consequently, for any $x \notin B$,

$$S_B(x) = \prod_{j=1}^d \vartheta_{\lambda,\varepsilon}(-\delta_j(x)) \ \leq \ \prod_{j=1}^d e^{-\lambda\,(\delta_j(x)-\varepsilon)_+} \ = \ \exp\Big(-\lambda \sum_{j=1}^d (\delta_j(x) - \varepsilon)_+\Big).$$

Since $\sum_{j=1}^d (\delta_j - \varepsilon)_+ \ \geq \ \big(\sum_j \delta_j - d\varepsilon\big)_+ = (Z - d\varepsilon)_+$, we have the pointwise bound

$$S_B(x) \ \leq \ \mathbf{1}\{Z \leq d\varepsilon\} \ + \ e^{-\lambda\,(Z - d\varepsilon)}\,\mathbf{1}\{Z > d\varepsilon\} \ \leq \ \mathbf{1}\{Z \leq d\varepsilon\} \ + \ e^{\lambda d\varepsilon}\,e^{-\lambda Z}.$$

Taking expectations and using $Z \geq Z_2$ (i.e., $d_1 \geq \operatorname{dist}$),

$$\mathbb{E}\big[S_B(X)\,\mathbf{1}_{B^c}(X)\big] \ \leq \ P\{Z_2 \leq d\varepsilon\} \ + \ e^{\lambda d\varepsilon}\,\mathbb{E}\big[e^{-\lambda Z}\big].$$

By the tube–mass hypothesis, $P\{Z_2 \leq d\varepsilon\} \leq C\,(d\varepsilon)^\beta = C\,(dc_0)^\beta\,\lambda^{-\beta}$.

For the Laplace term, the splitting bound

$$\mathbb{E}[e^{-\lambda Z}] \ \leq \ \mathsf{T}(\tau) + e^{-\lambda\tau} \qquad (\tau > 0)$$

combined with $\mathsf{T}(t) \leq Ct^\beta$ and the choice $\tau = \beta/\lambda$ yields

$$\mathbb{E}[e^{-\lambda Z}] \ \leq \ C\big(\tfrac{\beta}{\lambda}\big)^\beta + e^{-\beta} \ \leq \ C'_\beta\,\lambda^{-\beta},$$

for some constant C'_β depending only on β (since $\lambda \geq 1$). Since $e^{\lambda d\varepsilon} = e^{dc_0}$, we obtain

$$\mathbb{E}\big[S_B(X)\,\mathbf{1}_{B^c}(X)\big] \ \leq \ C\,(dc_0)^\beta\,\lambda^{-\beta} \ + \ e^{dc_0}\,C'_\beta\,\lambda^{-\beta},$$

which gives the stated $O(\lambda^{-\beta})$ bound.

■

E.1 Proof of upper bound on the $\mathcal{R}$TV

Now, we show the proof of the upper bound on the Radon total-variation as stated in Theorem 9 for a single box from Section 6.

First, we estblish some auxiliary lemmas about the 1D barrier function $\vartheta_{\lambda,\varepsilon}$, and its use in constructing box indicators with controlled derivatives.

Lemma 15 (Monotone C^{d+1} barrier with exact plateau) *Fix integers $d \geq 1$, $\lambda \geq 1$, and $\varepsilon \in (0,1]$. Let $H \in C^\infty(\mathbb{R})$ be nondecreasing with*

$$H(s) = 0 \ (s \leq 0), \qquad H(s) = 1 \ (s \geq 1), \qquad H^{(m)}(0) = H^{(m)}(1) = 0 \ \ (1 \leq m \leq d{+}1).$$

Define the scaled step $h_\varepsilon(t) := H\big((t+\varepsilon)/\varepsilon\big)$ (so $h_\varepsilon = 0$ on $(-\infty, -\varepsilon]$, $h_\varepsilon = 1$ on $[0,\infty)$, and $h'_\varepsilon \geq 0$ supported in $[-\varepsilon, 0]$), and set

$$\vartheta_{\lambda,\varepsilon}(t) \ := \ (1 - h_\varepsilon(t))\, e^{\lambda t} \ + \ h_\varepsilon(t).$$

Then $\vartheta_{\lambda,\varepsilon} \in C^{d+1}(\mathbb{R})$, is nondecreasing, and

$$\vartheta_{\lambda,\varepsilon}(t) = e^{\lambda t} \quad (t \leq -\varepsilon), \qquad \vartheta_{\lambda,\varepsilon}(t) = 1 \quad (t \geq 0).$$

For every $q \in \{1, \dots, d{+}1\}$ there exist constants C_q (depending only on q and H) such that

$$\|\vartheta^{(q)}_{\lambda,\varepsilon}\|_{L^\infty(\mathbb{R})} \ \leq \ C_q \sum_{m=0}^{q} \lambda^{\,q-m}\, \varepsilon^{-m}, \qquad \|\vartheta^{(q)}_{\lambda,\varepsilon}\|_{L^1(\mathbb{R})} \ \leq \ C_q \Big(\lambda^{q-1} + \sum_{m=1}^{q} \lambda^{\,q-m}\, \varepsilon^{\,1-m} \Big).$$

In particular, for $\varepsilon = c_0/\lambda$ (fixed $c_0 > 0$),

$$\|\vartheta^{(q)}_{\lambda,\varepsilon}\|_{L^\infty} \ \leq \ C'_q\, \lambda^q, \qquad \|\vartheta^{(q)}_{\lambda,\varepsilon}\|_{L^1} \ \leq \ C'_q\, \lambda^{q-1}.$$

Proof Monotonicity on $[-\varepsilon, 0]$ follows from $\vartheta' = (1 - h_\varepsilon)\lambda e^{\lambda t} + h'_\varepsilon(1 - e^{\lambda t}) \geq 0$, since $h'_\varepsilon \geq 0$ and $1 - e^{\lambda t} \geq 0$ for $t \leq 0$; outside, ϑ is $e^{\lambda t}$ (increasing) or 1. Leibniz and the scaling $\|h^{(m)}_\varepsilon\|_\infty \leq C_m \varepsilon^{-m}$, $\|h^{(m)}_\varepsilon\|_{L^1} \leq C_m \varepsilon^{1-m}$, plus $\int_{(-\infty,-\varepsilon]} |\partial_t^q e^{\lambda t}| dt = \lambda^{q-1} e^{-\lambda\varepsilon} \leq \lambda^{q-1}$, give the bounds. ■

Lemma 16 (1D product with sharp L^1 bounds) *Let $U(t) = \vartheta_{\lambda,\varepsilon}(u - t)$, $L(t) = \vartheta_{\lambda,\varepsilon}(t - \ell)$ with $u > \ell$. For $q \geq 0$ set $F_q := \partial_t^q(UL)$. Then*

$$\|F_0\|_{L^1(\mathbb{R})} \ \leq \ (u - \ell) \ + \ C\left(\varepsilon + \tfrac{1}{\lambda}\right), \qquad \|F_q\|_{L^1(\mathbb{R})} \ \leq \ C_q\, \lambda^{\,q-1} \quad (q \geq 1).$$

Proof Split $\mathbb{R}$ into $[\ell, u]$ (mass $u - \ell$), the two transition layers $[\ell - \varepsilon, \ell] \cup [u, u + \varepsilon]$ (mass $\leq 2\varepsilon$), and the tails where U or L is $\leq e^{-\lambda \cdot \mathrm{dist}}$ (mass $\leq 2/\lambda$). For $q \geq 1$, use Leibniz and place one factor in L^1 (Lemma 15) and the other in L^∞; when the undifferentiated factor appears, keep it in L^∞. ■

Lemma 17 (Per-box L^p bounds) *Assume $\varepsilon = c_0/\lambda$ with fixed $c_0 > 0$. Let $B = \prod_{j=1}^d [\ell_j, u_j]$ and $S_B(x) = \prod_{j=1}^d G_j(x_j)$ with $G_j(x_j) = \vartheta_{\lambda,\varepsilon}(u_j - x_j)\, \vartheta_{\lambda,\varepsilon}(x_j - \ell_j)$. For $1 \leq p \leq \infty$ and any multiindex α with $|\alpha| = s \geq 1$, write $q_j := \alpha_j$, $J := \{j : q_j \geq 1\}$, $r := |J|$. Then*

$$\|\partial^\alpha S_B\|_{L^p(\mathbb{R}^d)} \ \leq \ C_{d,s,p}\, \lambda^{\,s - \frac{r}{p}} \prod_{k \notin J} \Big((u_k - \ell_k) + \tfrac{C}{\lambda} \Big)^{\frac{1}{p}}.$$

In particular, $\|\partial^\alpha S_B\|_{L^\infty} \leq C_{d,s} \lambda^s$ and $\|\partial^\alpha S_B\|_{L^1} \leq C_{d,s} \lambda^{s-r} \prod_{k \notin J} \big((u_k - \ell_k) + \frac{C}{\lambda} \big)$.

Proof By separability, $\partial^\alpha S_B(x) = \prod_{j=1}^d \partial_{x_j}^{q_j} G_j(x_j)$ and, by Tonelli, $\|\partial^\alpha S_B\|_{L^p}^p = \prod_{j=1}^d \|\partial^{q_j} G_j\|_{L^p(\mathbb{R})}^p$.

For $q_j \geq 1$, Lemma 15 and Leibniz give $\|\partial^{q_j} G_j\|_{L^\infty} \lesssim \lambda^{q_j}$, and Lemma 16 gives $\|\partial^{q_j} G_j\|_{L^1} \lesssim \lambda^{q_j - 1}$. Interpolating, $\|\partial^{q_j} G_j\|_{L^p} \leq \|\cdot\|_\infty^{1-1/p} \|\cdot\|_1^{1/p} \lesssim \lambda^{q_j - \frac{1}{p}}$.

For $q_j = 0$, $\|G_j\|_{L^p} \leq \|G_j\|_\infty^{1-1/p} \|G_j\|_1^{1/p} \leq \big((u_j - \ell_j) + C/\lambda\big)^{1/p}$. Multiply over j. ■

Lemma 18 (Single-box aggregator bound) *Fix $\lambda \geq 1$ and $\varepsilon = c_0/\lambda$ with $c_0 > 0$. Let $B = \prod_{j=1}^d [\ell_j, u_j]$ and*

$$S_B(x) \;:=\; \prod_{j=1}^d G_j(x_j), \qquad G_j(x_j) := \vartheta_{\lambda,\varepsilon}(u_j - x_j)\, \vartheta_{\lambda,\varepsilon}(x_j - \ell_j).$$

Then

$$\sum_{|\alpha|=d+1} \|\partial^\alpha S_B\|_{L^1(\mathbb{R}^d)} \;\leq\; C_d \sum_{r=1}^d \lambda^{d+1-r}\, \mathcal{H}^{d-r}\big(\Sigma_{d-r}(B)\big),$$

where C_d depends only on d and H.

Proof

Step 1: Factorization by separability. Write $\alpha = (\alpha_1, \ldots, \alpha_d)$ with $|\alpha| = d+1$, and let $J := \{j : \alpha_j \geq 1\}$ (the active axes), $r := |J| \in \{1, \ldots, d\}$. By separability and Tonelli,

$$\partial^\alpha S_B(x) \;=\; \prod_{j=1}^d \partial_{x_j}^{\alpha_j} G_j(x_j), \qquad \|\partial^\alpha S_B\|_{L^1} \;=\; \prod_{j=1}^d \big\|\partial_{x_j}^{\alpha_j} G_j\big\|_{L^1(\mathbb{R})}.$$

Step 2: One-dimensional L^1 bounds. From Lemma 15 and Lemma 16 (with $\varepsilon = c_0/\lambda$),

$$\big\|\partial_{x_j}^q G_j\big\|_{L^1(\mathbb{R})} \;\leq\; \begin{cases} C\,\lambda^{q-1}, & q \geq 1, \\ (u_j - \ell_j) + C/\lambda, & q = 0. \end{cases}$$

Thus, for the multiindex α with active set J,

$$\|\partial^\alpha S_B\|_{L^1} \;\leq\; C^r\, \lambda^{\sum_{j\in J}(\alpha_j - 1)} \prod_{k \notin J} \Big((u_k - \ell_k) + \tfrac{C}{\lambda}\Big) \;=\; C^r\, \lambda^{d+1-r} \prod_{k \notin J} \Big((u_k - \ell_k) + \tfrac{C}{\lambda}\Big), \quad (33)$$

since $\sum_{j\in J} \alpha_j = d + 1$.

Step 3: Expanding the $(u_k - \ell_k) + C/\lambda$ factors. Let $a_k := u_k - \ell_k$ and $\beta := C/\lambda$. For fixed J,

$$\prod_{k \notin J} (a_k + \beta) \;=\; \sum_{L \subseteq J^c} \beta^{|L|} \prod_{k \notin J \cup L} a_k.$$

Insert this into (33):

$$\|\partial^\alpha S_B\|_{L^1} \;\leq\; C^r \sum_{L \subseteq J^c} \lambda^{d+1-r}\, \beta^{|L|} \prod_{k \notin J\cup L} a_k \;=\; C^r \sum_{L \subseteq J^c} \lambda^{d+1-(r+|L|)} \prod_{k \notin J \cup L} a_k.$$

Define $r' := r + |L| \in \{r, \ldots, d\}$. Grouping by r',

$$\|\partial^\alpha S_B\|_{L^1} \;\leq\; \sum_{r'=r}^d C_d\, \lambda^{d+1-r'} \sum_{\substack{L \subseteq J^c \\ |L| = r' - r}} \prod_{k \notin J \cup L} a_k. \quad (34)$$

Step 4: Summing over multiindices with the same active set J. For a fixed J with $|J| = r$, the number of compositions of $d{+}1$ into r strictly positive parts $(\alpha_j)_{j\in J}$ is $\binom{d}{r-1}$; absorbing this (and the C^r) into C_d, the sum over all α with $\operatorname{supp}(\alpha) = J$ yields

$$\sum_{\substack{\alpha:\ |\alpha|=d+1\\ \operatorname{supp}(\alpha)=J}} \|\partial^\alpha S_B\|_{L^1} \ \le\ \sum_{r'=r}^{d} C_d\, \lambda^{\,d+1-r'} \sum_{\substack{L\subseteq J^c\\ |L|=r'-r}} \prod_{k\notin J\cup L} a_k.$$

Step 5: Summing over active-axis choices J and identifying skeleton measures. Now sum over all $J \subseteq \{1,\dots,d\}$ with $|J| = r$, and then over $r = 1,\dots,d$. For a fixed r', the inner product $\prod_{k\notin J\cup L} a_k$ depends only on the union $J' := J \cup L$ with $|J'| = r'$; each such J' arises from finitely many pairs (J, L), which is absorbed into C_d. Hence

$$\sum_{\substack{J\subset[d]\\ |J|=r}} \sum_{\substack{L\subseteq J^c\\ |L|=r'-r}} \prod_{k\notin J\cup L} a_k \ \le\ C_d \sum_{\substack{J'\subset[d]\\ |J'|=r'}} \prod_{k\notin J'} a_k.$$

Recall that the $(d - r')$-skeleton measure of an axis-aligned box satisfies

$$\mathcal{H}^{d-r'}\big(\Sigma_{d-r'}(B)\big) \ =\ \sum_{\substack{J'\subset[d]\\ |J'|=r'}} 2^{r'} \prod_{k\notin J'} a_k,$$

because choosing J' fixes which r' coordinates are clamped to a face (each with two choices, ℓ or u), and the remaining coordinates span intervals of lengths a_k; overlaps of distinct faces have strictly lower dimension and therefore zero $\mathcal{H}^{d-r'}$-measure. Therefore,

$$\sum_{\substack{J'\subset[d]\\ |J'|=r'}} \prod_{k\notin J'} a_k \ =\ 2^{-r'}\,\mathcal{H}^{d-r'}\big(\Sigma_{d-r'}(B)\big) \ \le\ C_d\,\mathcal{H}^{d-r'}\big(\Sigma_{d-r'}(B)\big),$$

absorbing $2^{-r'}$ into C_d.

Step 6: Final aggregation. Putting Steps 4–5 together and summing $r' = 1,\dots,d$,

$$\sum_{|\alpha|=d+1} \|\partial^\alpha S_B\|_{L^1} \ \le\ \sum_{r'=1}^{d} C_d\,\lambda^{\,d+1-r'}\,\mathcal{H}^{d-r'}\big(\Sigma_{d-r'}(B)\big),$$

as claimed. ∎

Lemma 19 (Radon–RTV master inequality) *If $f \in C^{d+1}(\mathbb{R}^d)$ and $\partial^\alpha f \in L^1(\mathbb{R}^d)$ for all $|\alpha| = d{+}1$, then*

$$|f|_{\mathrm{RTV}} := \int_{S^{d-1}} \int_{\mathbb{R}} \big|\partial_t^{d+1}(\mathcal{R}f)(\beta,t)\big|\,dt\,d\beta \ \le\ C_d \sum_{|\alpha|=d+1} \|\partial^\alpha f\|_{L^1(\mathbb{R}^d)}.$$

Proof For fixed β, dominated convergence (majorant: a fixed linear combination of $|\partial^\alpha f|$, $|\alpha| = d{+}1$) gives $\partial_t^{d+1}(\mathcal{R}f)(\beta,t) = \int_{\beta^\perp} (\beta\cdot\nabla)^{d+1} f(y+t\beta)\, d\mathcal{H}^{d-1}(y)$. Integrate in t and apply Fubini to get $\int_{\mathbb{R}} |\partial_t^{d+1}(\mathcal{R}f)| \le \int_{\mathbb{R}^d} |(\beta\cdot\nabla)^{d+1} f|$, then average over $\beta \in S^{d-1}$ and bound $(\beta\cdot\nabla)^{d+1}$ by a linear combination of Cartesian derivatives. ■

Theorem 20 (Bounded RTV score for a single box) *With S_B as in Lemma 18 and $\varepsilon = c_0/\lambda$,*

$$\|S_B\|_{\mathcal{R}\mathrm{TV}} \;\le\; C_d \sum_{r=1}^{d} \lambda^{\,d+1-r}\, \mathcal{H}^{d-r}\big(\Sigma_{d-r}(B)\big).$$

Proof Apply Lemma 19 to $f = S_B$ and Lemma 18. ■

New approach