What functions does XGBoost learn?

Dohyeong Ki

Department of Statistics, UC Berkeley

Jan 20, 2026

Joint work with Aditya Guntuboyina; Available at https://arxiv.org/abs/2601.05444

XGBoost

Why XGBoost?

It is one of the most widely used off-the-shelf machine learning methods.

[PDF] **XGBoost**: A Scalable Tree Boosting System
T Chen - Cornell University, 2016 - medial-earlysign.github.io
XGBoost: A Scalable Tree Boosting System **XGBoost**: A Scalable Tree Boosting System
Save Cite Cited by 70712 Related articles

For tabular data,

XGBoost is consistently reported as a state-of-the-art method

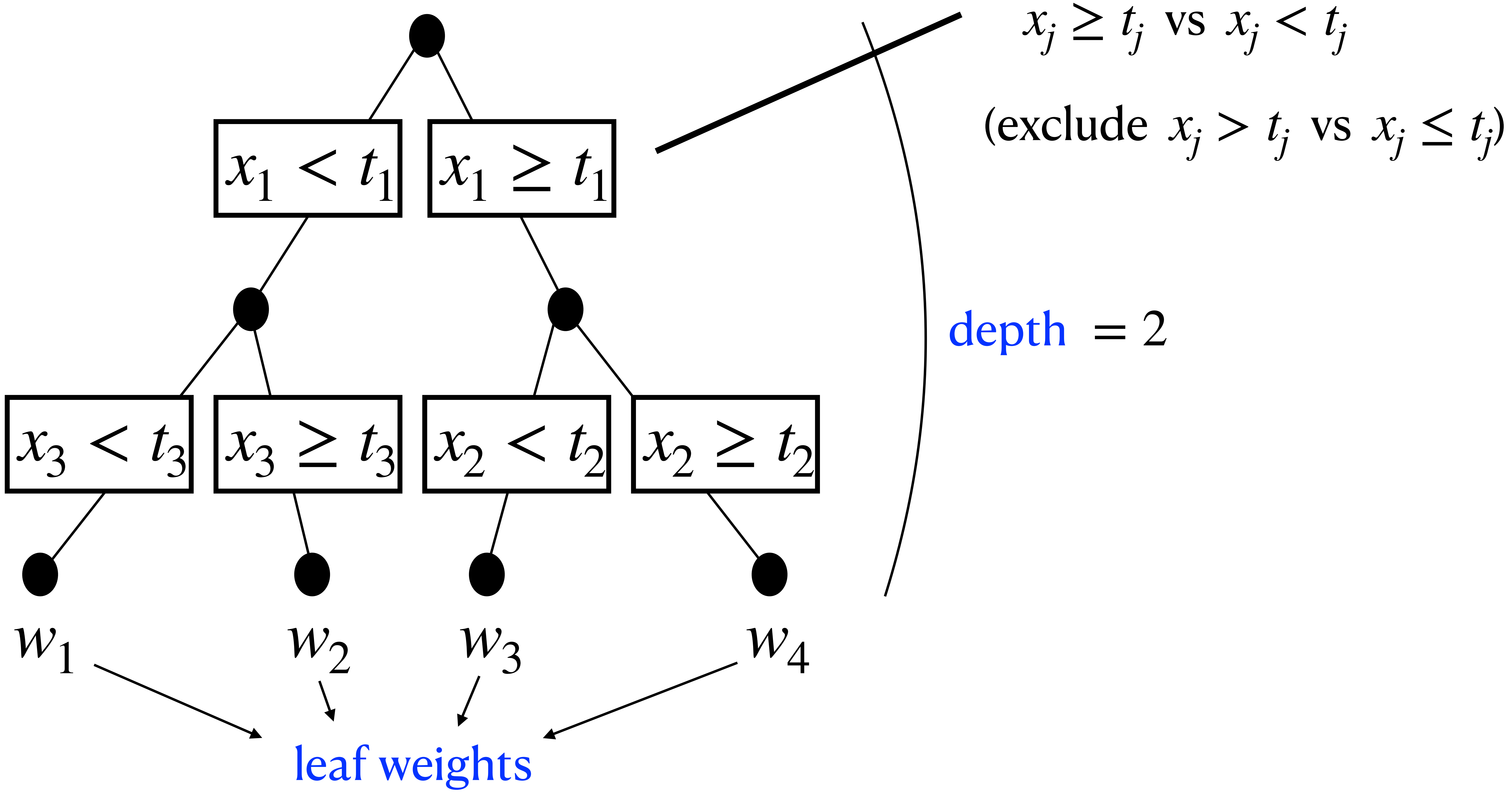
and often outperforms deep learning models;

see, e.g.,

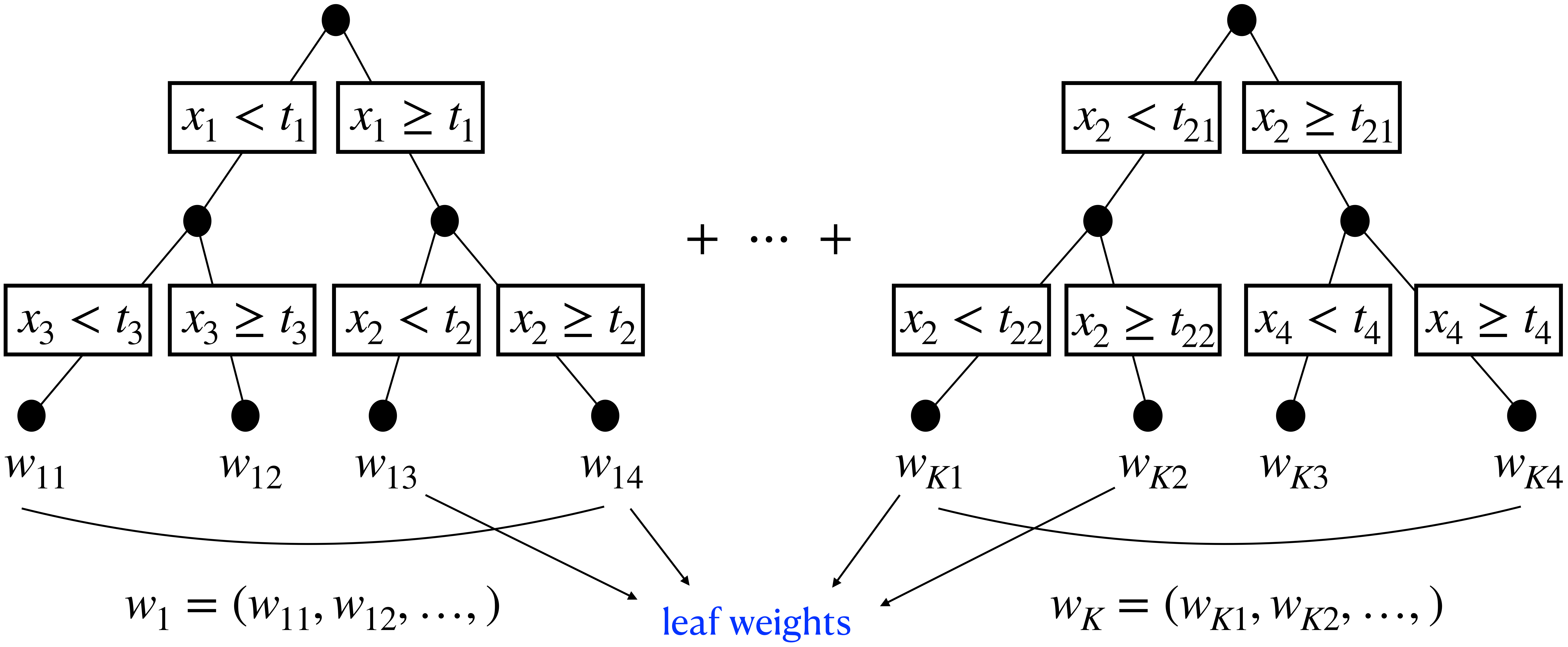
[Borisov et al. 22], [Grinsztajn, Oyallon, Varoquaux 22], [Shwartz-Ziv, Armon 22]

XGBoost fits a finite sum of regression trees to data.

Regression tree?



XGBoost fits a finite sum of regression trees to data.



XGBoost Optimization Problem

Given $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$ $(\mathbf{x}^{(i)} \in \mathbb{R}^d, y_i \in \mathbb{R})$, XGBoost aims to minimize

$$\sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 + \alpha \sum_k \|w_k\|_1$$

(1) squared L^2 norm is also common

(2) leaf-counting penalty $\gamma \sum T_k$ can also be imposed, where T_k is the number of leaves in the kth tree

over finite sums of regression trees with depth $\leq s$,

where w_k is the leaf weight vector of the kth tree.

→ XGBoost solves this problem using its iterative and greedy algorithm.

XGBoost Iterative Algorithm

Suppose after iteration $k-1$,

the fitted function is $\hat{f}^{(k-1)}$ (the sum of $k-1$ regression trees).

At iteration k, XGBoost optimizes the following in a greedy way:

$$\hat{f}_k \in \mathrm{argmin}_{f_k:\text{tree}} \left\{ \sum_{i=1}^{n} \Big(\underbrace{y_i - \hat{f}^{(k-1)}(\mathbf{x}^{(i)})}_{\text{current residuals}} - f_k(\mathbf{x}^{(i)}) \Big)^2 + \alpha \|w_k\|_1 \right\}$$

Update $\hat{f}^{(k-1)}$ to $\hat{f}^{(k)} = \hat{f}^{(k-1)} + \eta\, \hat{f}_k$, where $\eta \in (0,1)$ is a learning rate.

Motivating Question

Despite its popularity, XGBoost is not well studied theoretically.

In particular, it is not well understood that

Q. What kinds of functions can be learned accurately by XGBoost?

Q. What function class is XGBoost implicitly targeting?

This work answers these questions (at least in part) by studying
the XGBoost optimization problem and its objective function and solution
(but not its iterative and greedy algorithm).

XGBoost Optimization Problem

Given $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$ $(\mathbf{x}^{(i)} \in \mathbb{R}^d, y_i \in \mathbb{R})$, XGBoost aims to minimize

$$\sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 + \alpha \sum_{k} \|w_k\|_1 \longrightarrow \text{depends on sum-of-trees representations}$$

over finite sums of regression trees with depth $\leq s$,

where w_k is the leaf weight vector of the kth tree.

For each finite sum of regression trees f, define

$$V_{\mathrm{XGB}}^{d,s}(f) = \inf \left\{ \sum_k \|w_k\|_1 \right\}$$

where the infimum is over all representations of f into a finite sum of trees.

Let $\mathscr{F}_{\mathrm{ST}}^{d,s}$ denote the class of finite sums of regression trees with depth $\leq s$.

We can write the XGBoost optimization problem as

$$\operatorname{argmin}\left\{ \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 + \alpha V_{\mathrm{XGB}}^{d,s}(f) : f \in \mathscr{F}_{\mathrm{ST}}^{d,s} \right\}.$$

Preview

Every solution to the XGBoost optimization problem

$$\operatorname{argmin}\left\{ \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 + \alpha V^{d,s}_{\mathrm{XGB}}(f) : f \in \mathscr{F}^{d,s}_{\mathrm{ST}} \right\}$$

is also a solution to

extensions

$$\operatorname{argmin}\left\{ \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 + \alpha V^{d,s}_{\infty\text{-}\mathrm{XGB}}(f) : f \in \mathscr{F}^{d,s}_{\infty\text{-}\mathrm{ST}} \right\}.$$

$\rightarrow$ XGBoost is implicitly targeting a larger function class $\mathscr{F}^{d,s}_{\infty\text{-}\mathrm{ST}}$.

We will construct this function class $\mathscr{F}^{d,s}_{\infty\text{-}\mathrm{ST}}$ along with $V^{d,s}_{\infty\text{-}\mathrm{XGB}}(\,\cdot\,)$.

Basis for Finite Sums of Regression Trees

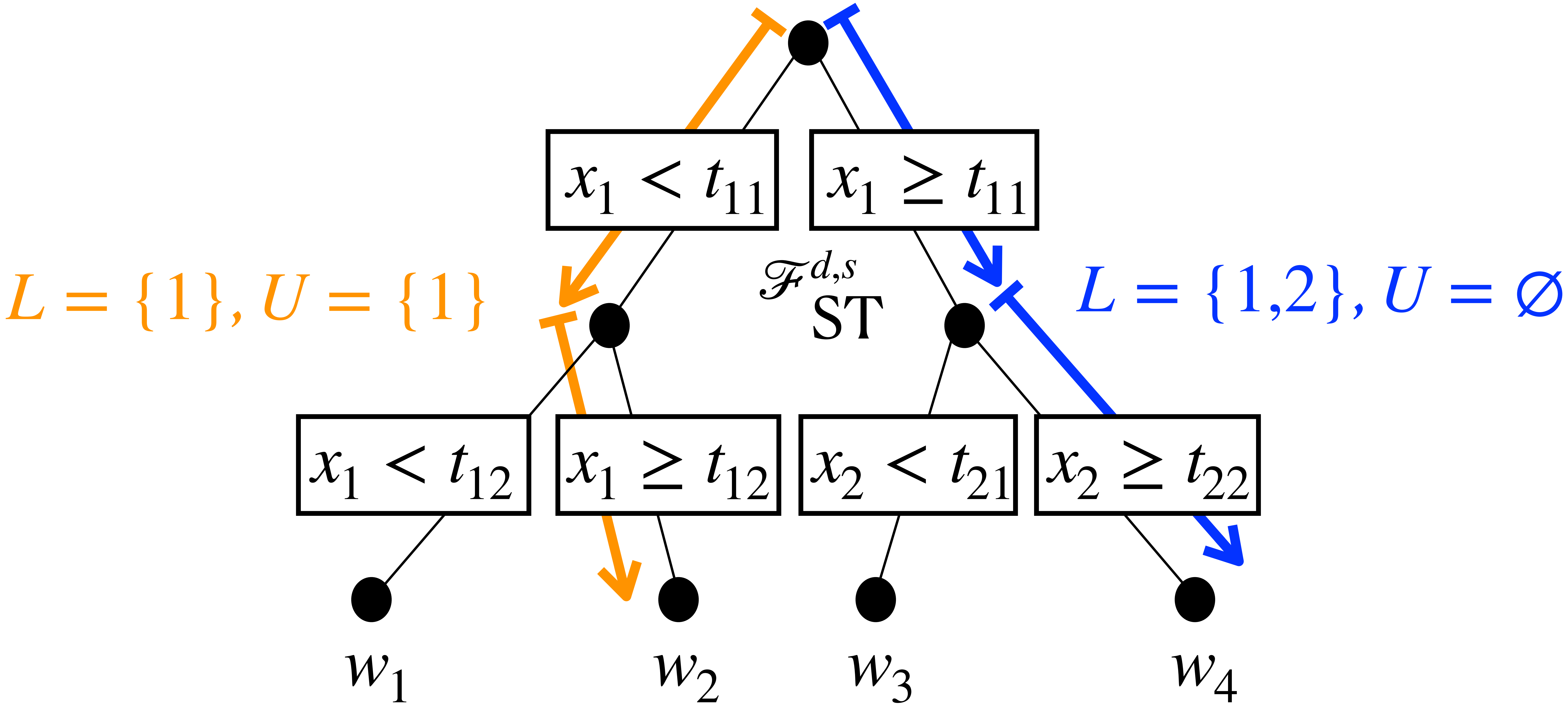
Every finite sum of regression trees with depth $\leq s$ (= every element of $\mathscr{F}^{d,s}_{\mathrm{ST}}$) can be expressed as a finite linear combination of

$$b^{L,U}_{\mathbf{l},\mathbf{u}}(x_1, \ldots, x_d) := \prod_{j \in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j \in U} \mathbf{1}(x_j < u_j)$$

where (1) $L, U \subseteq \{1, \ldots, d\}$ (possibly empty and not necessarily disjoint)

(2) $|L| + |U| \leq s$, and (3) each $l_j, u_j \in \mathbb{R}$.

$$b_{\mathbf{l},\mathbf{u}}^{L,U}(x_1, \ldots, x_d) = \prod_{j \in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j \in U} \mathbf{1}(x_j < u_j)$$



$\mathscr{F}_{\mathrm{ST}}^{d,s}$ is the collection of finite linear combinations of $b_{\mathbf{l},\mathbf{u}}^{L,U}$ with $|L| + |U| \leq s$.

Infinite-Dimensional Extension

We consider infinite linear combinations of $b_{\mathbf{l},\mathbf{u}}^{L,U}$ with $|L| + |U| \leq s$.

We define $\mathscr{F}_{\infty-\mathrm{ST}}^{d,s}$ as the collection of all functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form:

$$f_{c,\{\nu_{L,U}\}}(x_1, \ldots, x_d) := c + \sum_{0<|L|+|U|\leq s} \int_{\mathbb{R}^{|L|+|U|}} b_{\mathbf{l},\mathbf{u}}^{L,U}(x_1, \ldots, x_d)\, d\nu_{L,U}(\mathbf{l}, \mathbf{u})$$

where $\nu_{L,U}$ are finite signed (Borel) measures on $\mathbb{R}^{|L|+|U|}$.

$\to \mathscr{F}_{\infty-\mathrm{ST}}^{d,s}$ is an infinite-dimensional extension of $\mathscr{F}_{\mathrm{ST}}^{d,s}$.

Complexity Measure

Define the complexity of $f \in \mathscr{F}^{d,s}_{\infty-\mathrm{ST}}$ as

$$V^{d,s}_{\infty-\mathrm{XGB}}(f) := \inf \left\{ \sum_{0<|L|+|U|\le s} \|\nu_{L,U}\|_{\mathrm{TV}} : f_{c,\{\nu_{L,U}\}} \equiv f \right\}$$

where the infimum is over all possible representations $f_{c,\{\nu_{L,U}\}}$ of f.

The total variation $\|\nu\|_{\mathrm{TV}}$ of a signed measure ν on $\mathbb{R}^m$ is given by

$$\|\nu\|_{\mathrm{TV}} = |\nu|(\mathbb{R}^m) = \sup_{\mathscr{P}:\text{partition of } \mathbb{R}^m} \sum_{P \in \mathscr{P}} |\nu(P)|.$$

Main Result 1:

If $f \in \mathcal{F}^{d,s}_{\mathrm{ST}}$, i.e., f is a finite sum of regression trees,

$$V^{d,s}_{\infty\text{-XGB}}(f) = V^{d,s}_{\mathrm{XGB}}(f) = \inf \left\{ \sum_k \|w_k\|_1 \right\}$$

where the infimum is over all representations of f into a finite sum of trees.

$\rightarrow V^{d,s}_{\infty\text{-XGB}}(\,\cdot\,)$ is an extension of the XGBoost penalty $V^{d,s}_{\mathrm{XGB}}(\,\cdot\,)$.

Main Result 2:

Every solution to

$$\operatorname{argmin}\left\{ \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 + \alpha V^{d,s}_{\mathrm{XGB}}(f) : f \in \mathscr{F}^{d,s}_{\mathrm{ST}} \right\}$$

extensions

is also a solution to

$$\operatorname{argmin}\left\{ \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 + \alpha V^{d,s}_{\infty\text{-}\mathrm{XGB}}(f) : f \in \mathscr{F}^{d,s}_{\infty\text{-}\mathrm{ST}} \right\}.$$

→ XGBoost is implicitly targeting a larger function class $\mathscr{F}^{d,s}_{\infty\text{-}\mathrm{ST}}$.

Smoothness Characterizations of $\mathscr{F}^{d,s}_{\infty-\mathbf{ST}}$ and $V^{d,s}_{\infty-\mathbf{XGB}}(\cdot)$

$V^{d,s}_{\infty-\mathrm{XGB}}(\cdot)$ is closely related to Hardy–Krause variation ([Hardy 1905], [Krause 1903], [Aistleitner and Dick 15], [Leonov 96], [Owen 05]).

Hardy–Krause variation has been used for non-parametric regression; e.g., in

[Fang, Guntuboyina, and Sen 21], $\longrightarrow$ Hardy–Krause variation denoising

[Benkeser and van der Laan 16], [Schuler, Li, and van der Laan 22],

[van der Laan, Benkeser, and Cai 23] $\longrightarrow$ Highly Adaptive Lasso

Hardy–Krause Variation ($d = 2$)

For sufficiently smooth function f,

mixed partial derivatives of max order 1

$$\mathrm{HK}(f) = \int_{\mathbb{R}^2} |\underline{f^{(1,1)}}(x_1, x_2)|\, dx_1 dx_2$$

$$+ \int_{\mathbb{R}} |\underline{f^{(1,0)}}(x_1, -\infty)|\, dx_1 + \int_{\mathbb{R}} |\underline{f^{(0,1)}}(-\infty, x_2)|\, dx_2 \,.$$

L^p norm constraints on mixed partial derivatives have been used for

nonparametric regression ([Fang, Guntuboyina, and Sen 21], [Lin 00], etc.)

approximation/interpolation

([Dũng, Temlyakov, and Ullrich 18], [Bungartz and Griebel 04], etc.)

Smoothness Characterization of $\mathscr{F}^{d,s}_{\infty\text{–}\mathbf{ST}}$

When $s = d$,

$$\mathscr{F}^{d,d}_{\infty\text{–ST}} = \left\{ f : \mathrm{HK}(f) < \infty \ \text{ and } \ f \ \text{ is right-continuous} \right\}.$$

When $s < d$, we need some extra condition.

For example, if $d = 2$ and $s = 1$, we need to add that

$$f(v_1, v_2) - f(u_1, v_2) - f(v_1, u_2) + f(u_1, u_2) = 0$$

for all $u_1 < v_1$ and $u_2 < v_2$.

Comparison of $V^{d,s}_{\infty\text{-XGB}}(\cdot)$ to Hardy–Krause Variation

For every $f \in \mathcal{F}^{d,s}_{\infty\text{-ST}}$,

$$\mathrm{HK}(f)/\min(2^s - 1, 2^d) \leq V^{d,s}_{\infty\text{-XGB}}(f) \leq \mathrm{HK}(f).$$

XGBoost has been regarded as a purely algorithmic method.

But these characterizations suggest that XGBoost can be viewed as

a smoothness-constrained nonparametric regression method.

Theoretical Accuracy

We study the theoretical accuracy via the constrained version:

$$\hat{f}^{d,s}_{n,V} \in \operatorname{argmin}\left\{ \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 : f \in \mathscr{F}^{d,s}_{\mathrm{ST}} \text{ and } V^{d,s}_{\mathrm{XGB}}(f) \leq V \right\}.$$

Assume the standard random design setting:

can be replaced by a weaker assumption

(1) $y_i = f^*(\mathbf{x}^{(i)}) + \epsilon_i$ where $f^* \in \mathscr{F}^{d,s}_{\infty\text{-ST}}$ and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\, \sigma^2)$

(2) $\mathbf{x}^{(i)} \overset{\text{i.i.d.}}{\sim} p_0$ for some density p_0 with compact support and bounded above.

Main Result 3:

If $V > V^{d,s}_{\infty\text{-XGB}}(f^*)$, then we have

constant factor depends on s, V, and σ

$$\mathbb{E}\left[\int \left(\hat{f}^{d,s}_{n,V}(\mathbf{x}) - f^*(\mathbf{x})\right)^2 \cdot p_0(\mathbf{x})\, d\mathbf{x}\right] = O\left(\mathrm{poly}(d) \cdot n^{-2/3}(\log n)^{4(\min(s,d)-1)/3}\right).$$

The nearly dimension-free rate $n^{-2/3}$ (with some log factor) indicates that

→ The XGBoost complexity $V^{d,s}_{\infty\text{-XGB}}(\cdot)$ (and $V^{d,s}_{\mathrm{XGB}}(\cdot)$) becomes proportionally more restrictive as the dimension d increases.

→ Elements of $\mathscr{F}^{d,s}_{\infty\text{-ST}}$ are expected to be learned accurately by XGBoost.

Summary

We study a natural infinite-dimensional function class, along with a complexity measure, for XGBoost

This function class sheds light on what functions XGBoost can learn efficiently

Complexity measure is closely related to Hardy–Krause variation

The solution to the XGBoost optimization problem achieves a nearly dimension-free rate of convergence

Whether XGBoost's algorithm achieves a similar rate is an open problem

References

Borisov, V. et al. (2022). Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems 35 (6), 7499–7519.

Grinsztajn, L., E. Oyallon, and G. Varoquaux (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems 35, 507–520.

Shwartz-Ziv, R. and A. Armon (2022). Tabular data: Deep learning is not all you need. Information Fusion 81, 84–90.

Hardy, G. H. (1905). On double Fourier series, and especially those which represent the double zeta-function with real and incommensurable parameters. Quarterly Journal of Mathematics 37, 53–79.

Krause, M. (1903). Über mittelwertsätze im gebiete der doppelsummen and doppelintegrale. Leipziger Ber. 55, 239–263.

Aistleitner, C. and J. Dick (2015). Functions of bounded variation, signed measures, and a general Koksma–Hlawka inequality. Acta Arithmetica 167 (2), 143–171.

Leonov, A. S. (1996). On the total variation for functions of several variables and a multidimensional analog of Helly's selection principle. Mathematical Notes 63 (1), 61–71.

Owen, A. B. (2005). Multidimensional variation for quasi-Monte Carlo. Contemporary Multivariate Analysis and Design of Experiments: In Celebration of Professor Kai-Tai Fang's 65th Birthday, 49–74.

Fang, B., A. Guntuboyina, and B. Sen (2021). Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. Ann. Statist. 49 (2), 769–792.

Benkeser, D. and M. van der Laan (2016). The highly adaptive lasso estimator. IEEE International Conference on Data Science and Advanced Analytics (DSAA), 689–696.

Schuler, A., Y. Li, and M. van der Laan (2022). Lassoed tree boosting. arXiv preprint arXiv:2205.10697.

van der Laan, M. J., D. Benkeser, and W. Cai (2023). Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. International Journal of Biostatistics 19 (1), 261–289.

Lin, Y. (2000). Tensor product space ANOVA models. Ann. Statist. 28 (3), 734–755.

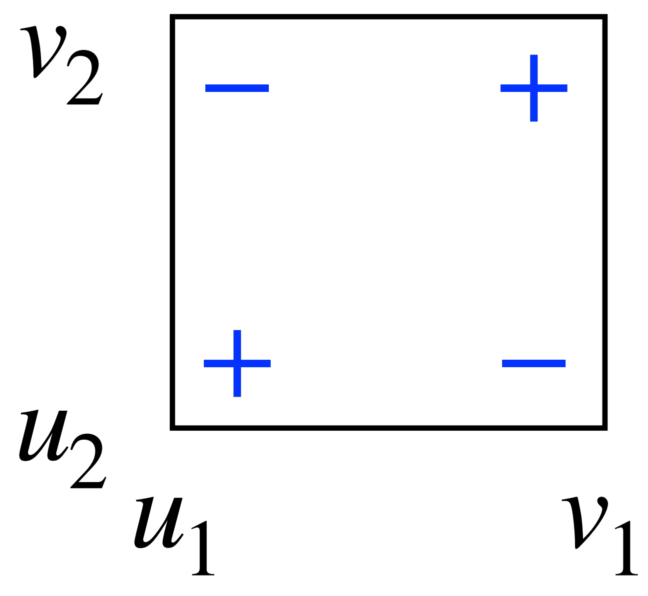
Dũng, D., V. Temlyakov, and T. Ullrich (2018). Hyperbolic Cross Approximation. Advanced Courses in Mathematics. CRM Barcelona. Birkhäuser, Cham.

Bungartz, H.-J. and M. Griebel (2004). Sparse grids. Acta Numerica 13, 147–269.

Friedman, J. H. (1991). Multivariate adaptive regression splines. Ann. Statist. 19 (1), 1–67.

Vitali Variation

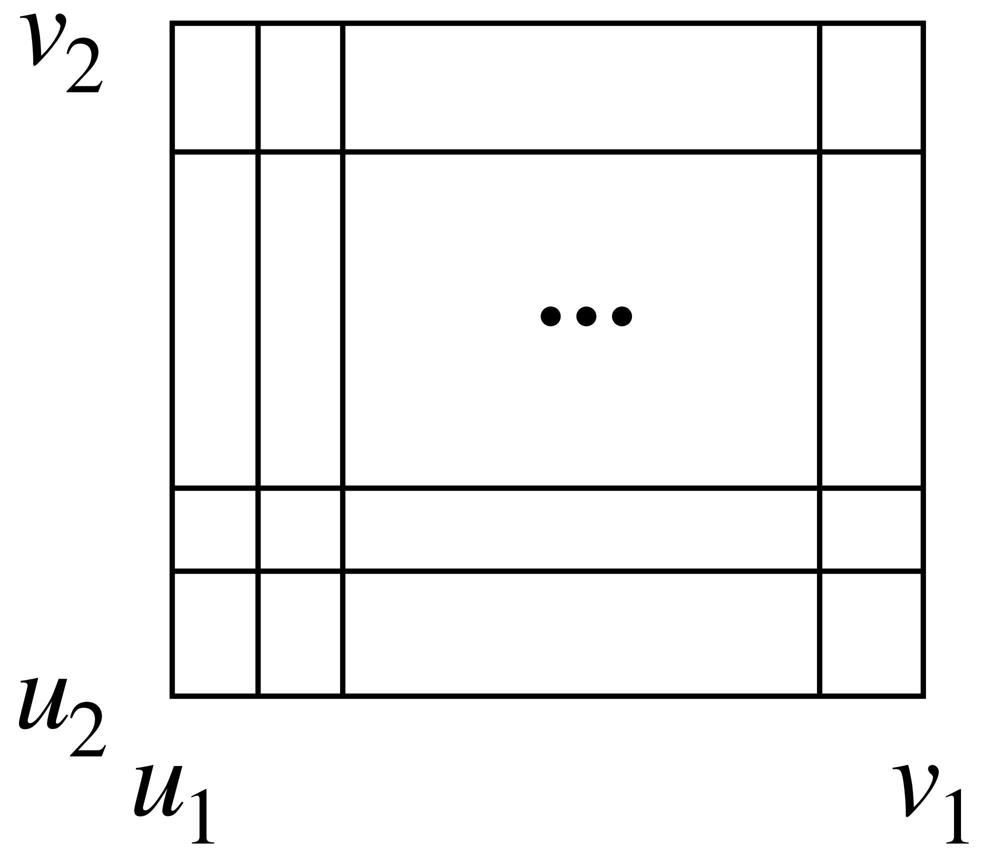
Let $g : \mathbb{R}^2 \to \mathbb{R}$. For $(u_1, u_2), (v_1, v_2) \in \mathbb{R}^2$ with $u_j < v_j$,



the quasi-volume of g on $[u_1, v_1] \times [u_2, v_2]$ is defined by

$$\Delta\big(g; [u_1, v_1] \times [u_2, v_2]\big) = g(v_1, v_2) - g(u_1, v_2) - g(v_1, u_2) + g(u_1, u_2)\,.$$

The Vitali variation of g on $[u_1, v_1] \times [u_2, v_2]$ is defined by



$$\mathrm{Vit}\big(g; [u_1, v_1] \times [u_2, v_2]\big) = \sup_{\mathscr{P}} \sum_{R \in \mathscr{P}} |\,\Delta(g; R)\,|$$

where the supremum is taken over all axis-aligned splits $\mathscr{P}$ of $[u_1, v_1] \times [u_2, v_2]$.

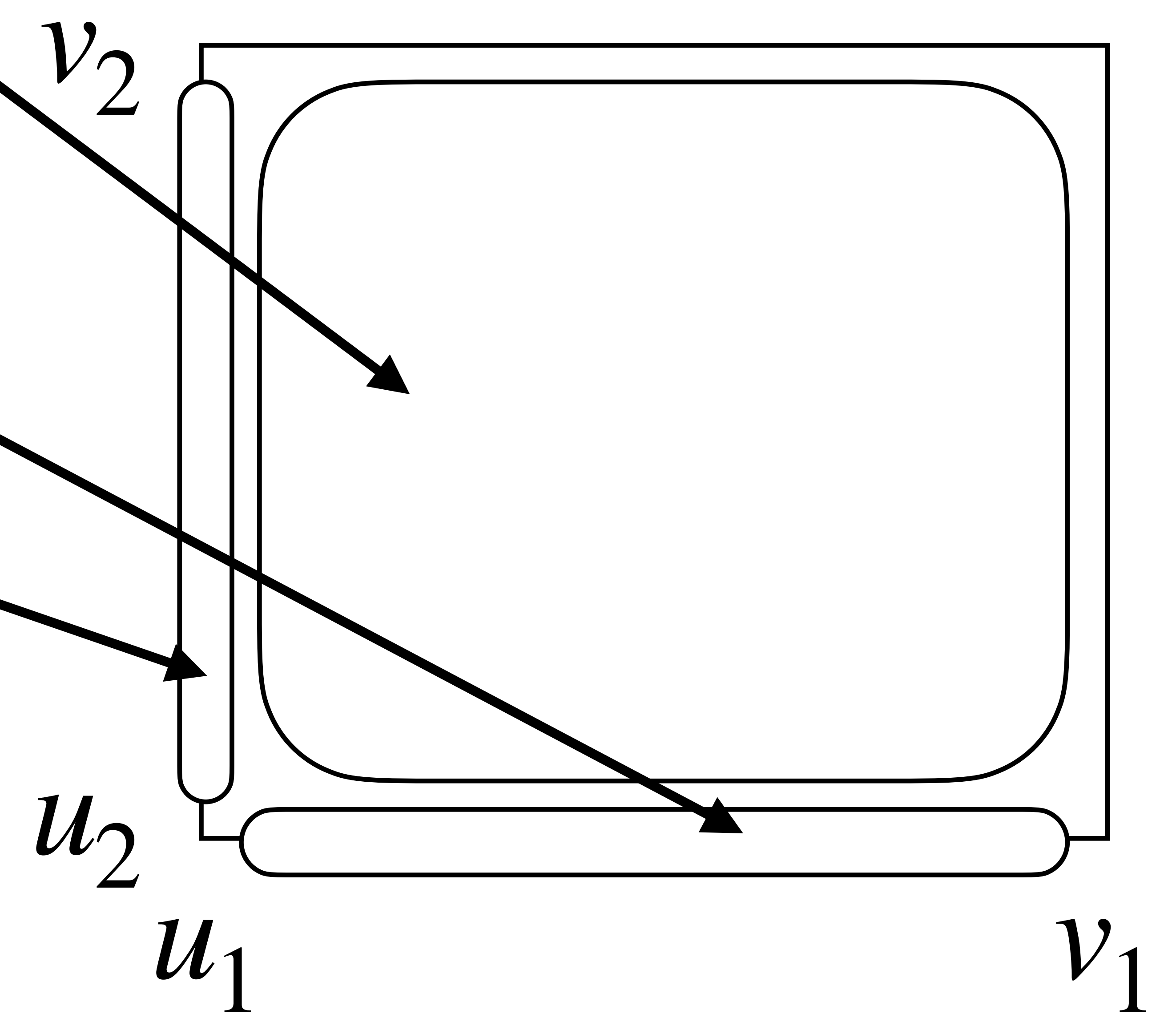
Hardy–Krause Variation on Compact Domains

Let $f : [u_1, v_1] \times [u_2, v_2] \to \mathbb{R}$.

any other corner of the domain
can be used for the anchor

The Hardy–Krause variation of f anchored at (u_1, u_2) is defined by

$$\begin{aligned} \mathrm{HK}\big(f; [u_1, v_1] \times [u_2, v_2]\big) = {} & \mathrm{Vit}\big(f; [u_1, v_1] \times [u_2, v_2]\big) \\ & + \mathrm{Vit}\big(x_1 \mapsto f(x_1, u_2); [u_1, v_1]\big) \\ & + \mathrm{Vit}\big(x_2 \mapsto f(u_1, x_2); [u_2, v_2]\big) \end{aligned}$$



Hardy–Krause Variation on $\mathbb{R}^2$

Let $f : \mathbb{R}^2 \to \mathbb{R}$.

The Hardy–Krause variation of f anchored at $(-\infty, -\infty)$ is defined by

$$\begin{aligned}\mathrm{HK}(f) = &\sup_{u_1<v_1, u_2<v_2} \mathrm{Vit}\big(f; [u_1, v_1] \times [u_2, v_2]\big) \\ &+ \sup_{u_1<v_1} \mathrm{Vit}\big(x_1 \mapsto f(x_1, -\infty); [u_1, v_1]\big) \\ &+ \sup_{u_2<v_2} \mathrm{Vit}\big(x_2 \mapsto f(-\infty, x_2); [u_2, v_2]\big)\end{aligned}$$