What functions does XGBoost learn?

Dohyeong Ki

Department of Statistics, UC Berkeley

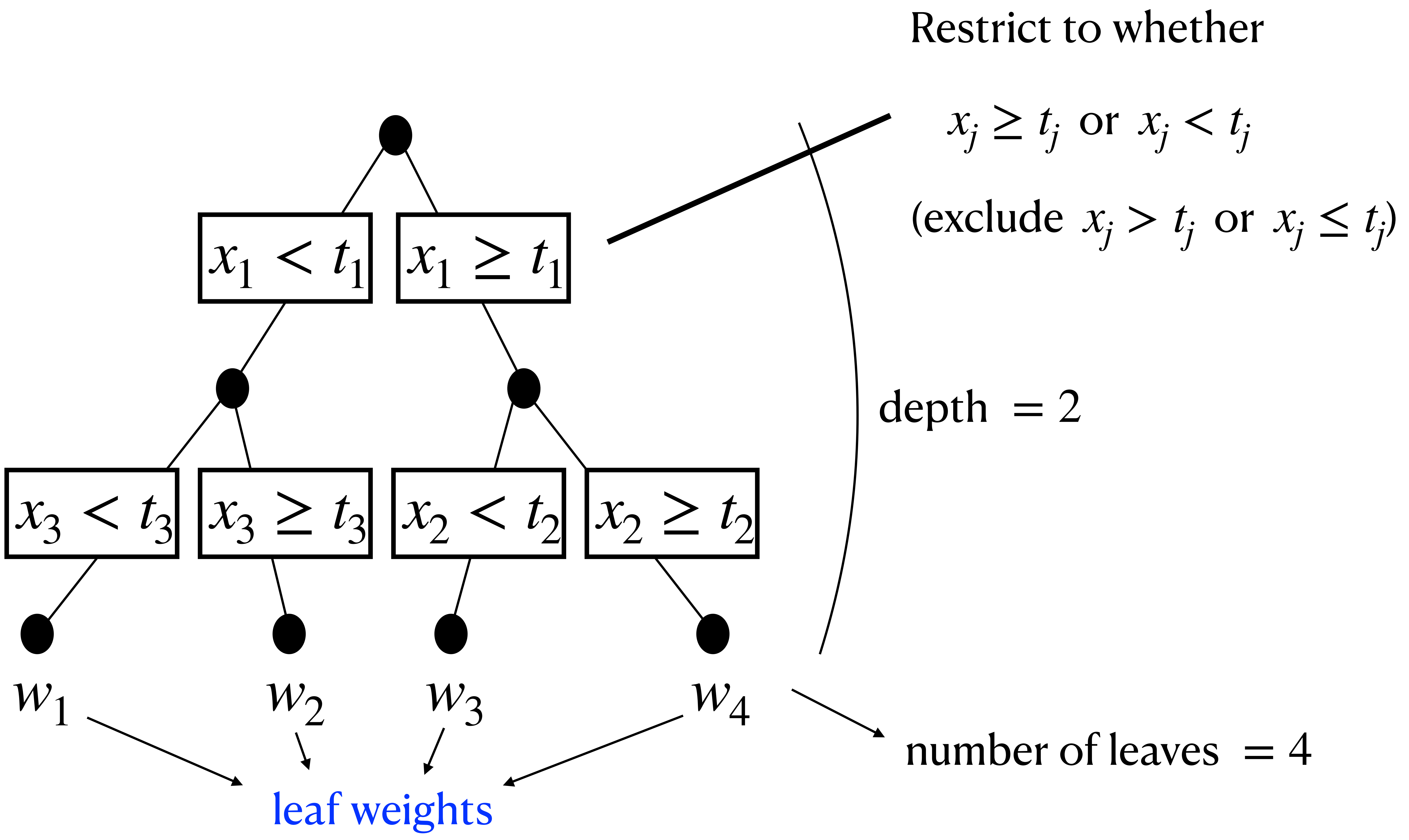
Dec 17, 2025

Joint work with Aditya Guntuboyina

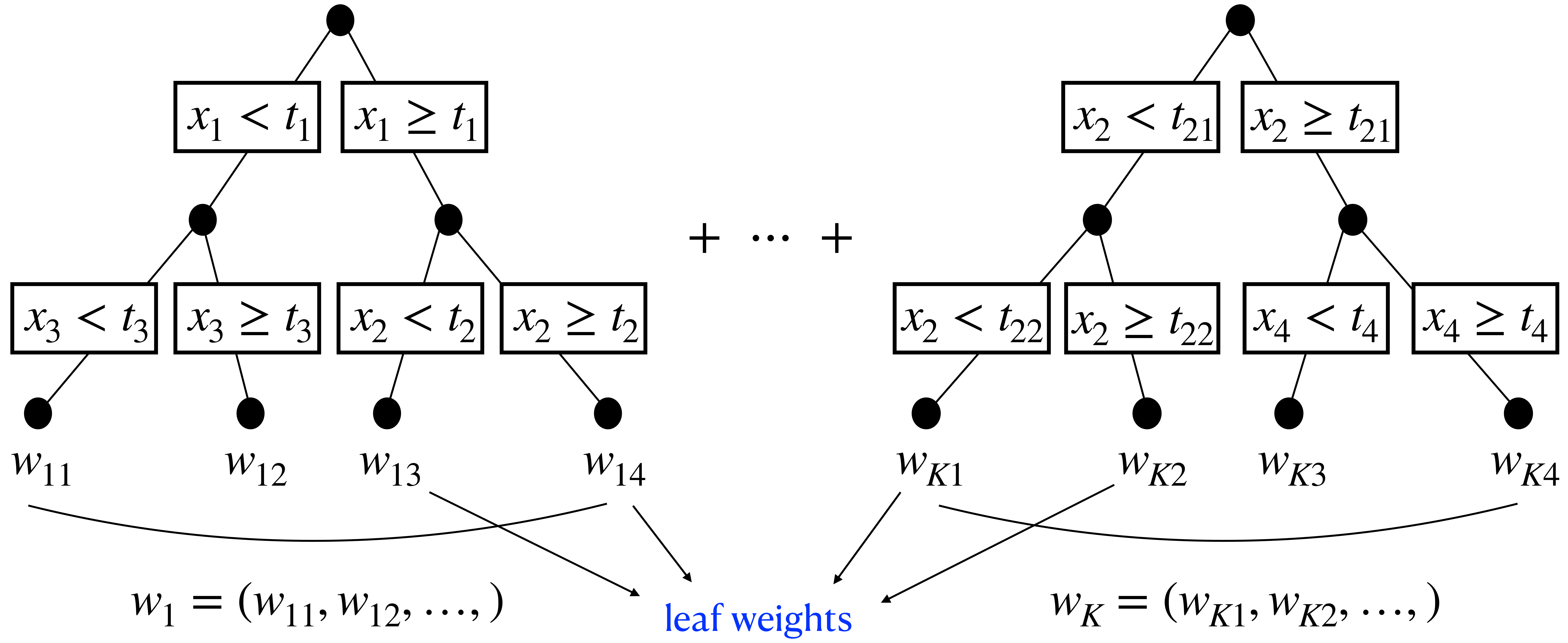
XGBoost

XGBoost fits a finite sum of regression trees to data.

Regression tree?



XGBoost fits a finite sum of regression trees to data.



Motivating Question

XGBoost produces a discrete-valued fit (it takes only finite different values), yet it seems to learn continuous functions quite well.

Q. What kinds of functions can XGBoost learn efficiently?

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

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

squared L^2 norm is also common

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

where (1) T_k is the number of leaves in the kth tree,

(2) w_k is its vector of leaf weights.

→ The solution to this problem can be seen as an idealized target of XGBoost

Q. What kinds of functions can XGBoost learn efficiently, in principle?

Finite Sums of Regression Trees

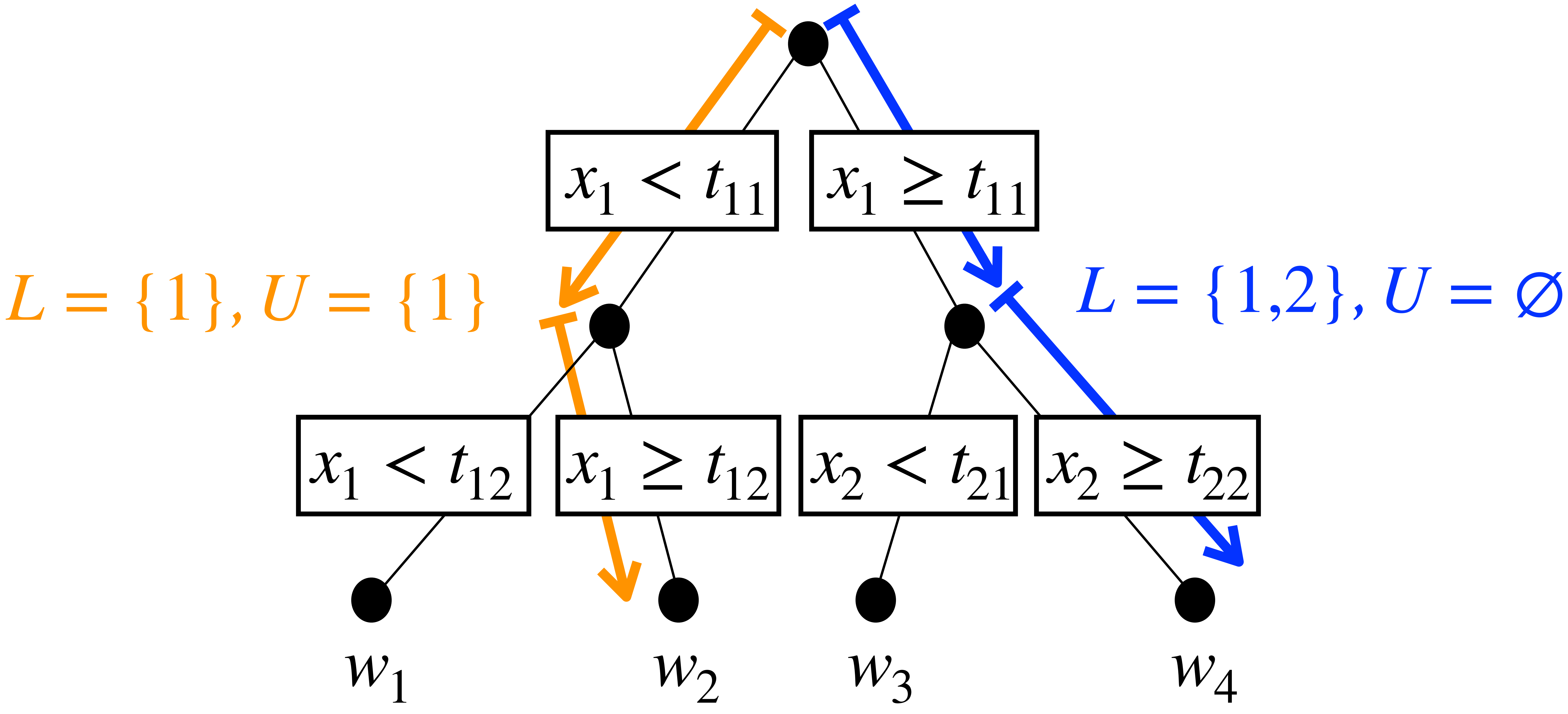
Every finite sum of regression trees with depth $\leq s$ can be expressed as a finite linear combination of

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

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

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

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



Infinite-Dimensional Extension

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

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

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

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

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

the class of finite sums of regression trees with depth $\leq s$.

Complexity Measure

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

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

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.

Recall that the XGBoost penalty is

$$\gamma \sum_k T_k + \alpha \sum_k \|w_k\|_1$$

Main Result 1:

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

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

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

$\rightarrow V^{d,s}_{\infty-\mathrm{XGB}}(\,\cdot\,)$ is an extension of the XGBoost penalty with $\gamma = 0$

$\gamma = 0$ means no penalty on numbers of leaves; the default choice by XGBoost

Idealized Target for XGBoost

Recall that we view

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

as an idealized target of XGBoost (with $\gamma = 0$).

The constrained version of this problem can be more formally written as

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

Main Result 2:

$\hat{f}^{d,s}_{n,V}$ is a least squares estimator over all $f \in \mathscr{F}^{d,s}_{\infty-\mathrm{ST}}$ with $V^{d,s}_{\infty-\mathrm{XGB}}(f) \leq V$.

→ Idealized target of XGBoost is, in fact, a solution to

the least squares problem over $\mathscr{F}^{d,s}_{\infty-\mathrm{ST}}$ with a constraint on $V^{d,s}_{\infty-\mathrm{XGB}}(\,\cdot\,)$

Further Insight into $\mathscr{F}^{d,s}_{\infty\text{–ST}}$ and $V^{d,s}_{\infty\text{–XGB}}(\cdot)$

$V^{d,s}_{\infty\text{–XGB}}(\cdot)$ is closely related to Hardy–Krause variation ([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], ⟶ Hardy–Krause variation denoising

[Benkeser and van der Laan 16],

[Schuler, Li, and van der Laan 22], ⟶ Highly Adaptive Lasso

[van der Laan, Benkeser, and Cai 23]

(1)

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

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

(2) For every $f \in \mathscr{F}^{d,s}_{\infty\text{-ST}}$,

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

Theoretical Accuracy of the Idealized Target

Assume the standard random design setting:

can be replaced by a weaker assumption

(1) $y_i = f^*(\mathbf{x}^{(i)}) + \epsilon_i$ where $f^* \in \mathscr{F}^{d,s}_{\infty-\mathrm{ST}}$ 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 that has compact support and is bounded above,

Main Result 3:

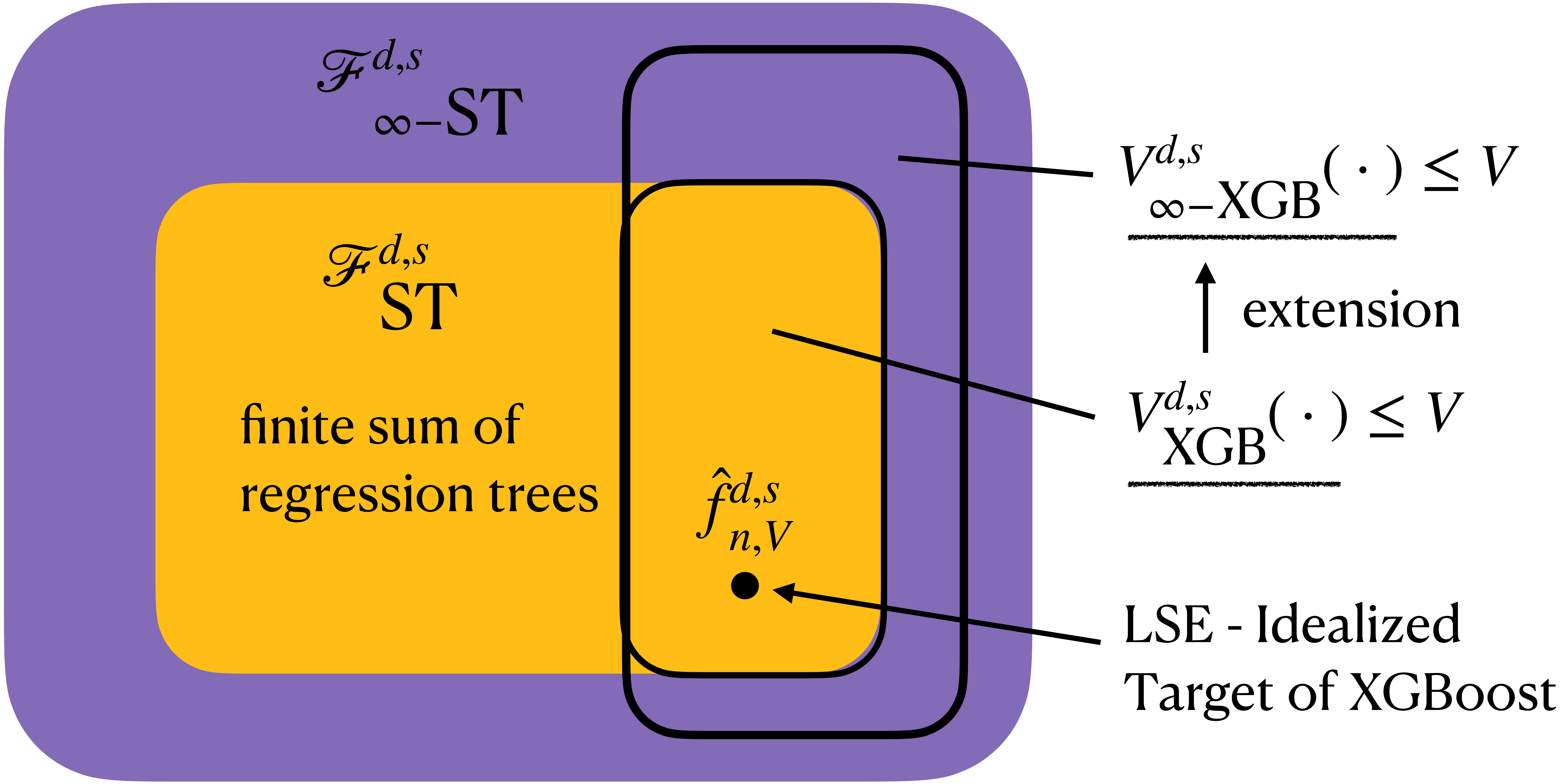
If $V > V^{d,s}_{\infty\text{-XGB}}(f^*)$, then we have

constant factor depends on s, V, and σ

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

This rate is also a nearly minimax optimal rate for the estimation over

$$\left\{f \in \mathcal{F}^{d,s}_{\infty\text{-ST}} : V^{d,s}_{\infty\text{-XGB}}(f) \le V\right\}.$$



Elements of $\mathcal{F}^{d,s}_{\infty-\mathrm{ST}}$ can be learned efficiently by XGBoost, in principle!

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 least squares estimator, which can be seen as an idealized target for XGBoost, achieves a nearly dimension-free rate of convergence

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

References

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.