Title

Machine Learning Methods Through the Lens of Infinite-Dimensional Lasso

ISBN

9798189270628

Author

Ki, Dohyeong

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Machine Learning Methods Through the Lens of Infinite-Dimensional Lasso

by

Dohyeong Ki

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Adityanand Guntuboyina, Chair
Professor Ryan Tibshirani
Assistant Professor Song Mei

Spring 2026

Machine Learning Methods Through the Lens of Infinite-Dimensional Lasso

Abstract

Machine Learning Methods Through the Lens of Infinite-Dimensional Lasso

by

Dohyeong Ki

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Adityanand Guntuboyina, Chair

We study a class of regression methods in machine learning that construct estimators as finite linear combinations of data-adaptively selected atoms, particularly multivariate adaptive regression splines (MARS) and one of the most widely used gradient boosting methods, XGBoost. Despite their popularity and empirical success, the theoretical properties of these methods remain poorly understood, largely due to the greedy procedures used in their estimator construction, which make rigorous analysis challenging.

We study these methods through a unifying infinite-dimensional optimization framework, in which functions are expressed as integrals of atoms with respect to finite signed measures. Within this framework, we consider a natural complexity measure given by the total variation of the representing measure, which serves as an infinite-dimensional analogue of the ℓ^1 norm of finite-dimensional coefficient vectors. This leads to an infinite-dimensional lasso formulation, which provides an idealized optimization perspective on these methods while abstracting away the greedy procedures used in practice.

We apply this framework to MARS and XGBoost. For MARS, it yields a variant of the original method that is amenable to theoretical analysis and can perform competitively in practice, albeit at higher computational cost. For XGBoost, it yields an optimization problem equivalent to the original formulation, but posed over a larger function class. This clarifies the function class implicitly underlying the method. In both settings, we show that the associated function classes and complexity measures admit characterizations in terms of smoothness. These results highlight a previously underappreciated connection between MARS, XGBoost, and smoothness-based nonparametric regression.

We also introduce a related shape-constrained regression method, termed totally concave regression, obtained by imposing sign constraints on the representing measure rather than the total variation constraint. It provides a multivariate extension of concave regression that differs from existing approaches based on classical notions of concavity. The proposed

method overcomes several limitations of classical approaches and offers a computationally and theoretically attractive alternative in multi-dimensional settings.

We study the statistical properties of the resulting constrained least squares estimators for all three methods—MARS, XGBoost, and totally concave regression. We show that our variant of MARS and totally concave regression can be computed via convex optimization, although they are formulated as infinite-dimensional optimization problems. Moreover, we establish rates of convergence for all three methods under standard regression models. In particular, we show that these estimators achieve nearly dimension-free rates, up to logarithmic factors. These results provide theoretical support for the effectiveness of these methods, complementing their strong empirical performance.

Contents

Chapter 1

Introduction

Given data $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$, where $\mathbf{x}^{(i)} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, a large class of regression methods in machine learning constructs estimators of the form

$$f(\mathbf{x}) = \sum_{k=1}^{K} c_k A_k(\mathbf{x}), \tag{1.1}$$

where c_k are real coefficients and A_k are real-valued functions—often referred to as atoms or basis functions—chosen in a data-adaptive manner. Prominent examples include multivariate adaptive regression splines (MARS) [53], boosting [50, 52], as well as random forests [19]. For MARS, the atoms are hinge functions of the form

$$(x_j - t)_+ := \max(0, x_j - t) \quad \text{or} \quad (t - x_j)_+ := \max(0, t - x_j)$$

where $t \in \mathbb{R}$ is referred to as a knot, together with products of such hinge functions. For tree-based boosting (e.g., [52, 29, 77, 123]) and random forests, the atoms are decision trees.

In many settings, the collection of candidate atoms is infinite (or finite but very large). For example, in tree-based boosting, the space of decision trees is infinite. While one may restrict split thresholds for decision trees to a finite set determined by the data, the resulting collection of trees remains extremely large. As a result, constructing an estimator of the form (1.1) requires selecting a finite subset of atoms from an infinite (or very large) collection, along with determining their coefficients. While the richness of the collection provides flexibility, it also poses computational challenges. Many machine learning methods address this through greedy procedures, such as stagewise selection or boosting, which iteratively select and add reasonably good (though not necessarily optimal) atoms to the estimator.

While such methods provide computationally tractable procedures for constructing estimators of the form (1.1), their theoretical properties are often not well understood. The use of greedy procedures makes the resulting estimators difficult to analyze, and a precise characterization of their statistical behavior remains elusive in many settings. In particular, it is often unclear what optimization problems these methods are implicitly solving, or whether the underlying objectives can be explicitly formulated. Moreover, there is limited understanding of the classes of functions to which these methods are well suited, and the conditions under which they are expected to perform well.

In this thesis, we take a step toward addressing these questions, focusing in particular on multivariate adaptive regression splines (MARS) [53] and one of the most widely used tree-based boosting methods, extreme gradient boosting (XGBoost) [29], by identifying and studying their underlying function classes and optimization problems. We study MARS in detail in Chapter 3 and XGBoost in Chapter 5.

Both methods construct estimators of the form (1.1), where the atoms A_k are hinge functions or their products for MARS, and decision trees for XGBoost. In both cases, the atoms can be indexed by a continuously varying parameter $\mathbf{t} \in D$ and can be written as $A_{\mathbf{t}}$ for some topological space D. For MARS, this parameter $\mathbf{t}$ corresponds to knot locations (see Chapter 3), while for XGBoost, it corresponds to split thresholds for decision trees (see Chapter 5). Using this notation, (1.1) can be written as

$$f(\mathbf{x}) = \sum_{k=1}^{K} c_k A_{\mathbf{t}_k}(\mathbf{x}) \tag{1.2}$$

for some $\mathbf{t}_1, \ldots, \mathbf{t}_K \in D$.

Although the estimator is always a (finite) linear combination of atoms as in (1.2), the fact that D can be infinite and that the locations of $\mathbf{t}_k$ are unknown makes analysis of these methods challenging. One way to circumvent this difficulty is to consider relaxations that allow for *infinite linear combinations* of atoms. In contrast to finite linear combinations, which are parameterized by finite-dimensional coefficient vectors, infinite linear combinations are represented using finite signed (Borel) measures. Specifically, for the atoms $A_{\mathbf{t}}$, such representations take the form

$$f_\nu(\mathbf{x}) := \int_D A_{\mathbf{t}}(\mathbf{x})\, d\nu(\mathbf{t}), \tag{1.3}$$

where ν is a finite signed (Borel) measure on D. This formulation can be viewed as a continuous generalization of finite linear combinations. Indeed, when ν is a discrete signed measure supported on $\mathbf{t}_1, \ldots, \mathbf{t}_K \in D$ with $\nu(\{\mathbf{t}_k\}) = c_k$, we recover (1.2). Since the coefficients c_k can be both positive and negative, here the measure ν should be a signed measure rather than a nonnegative measure.

Let $\mathcal{F}$ denote the class of all such infinite linear combinations (1.3). A natural estimator over this class $\mathcal{F}$ is the least squares estimator, defined as

$$\operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}^{(i)})\big)^2.$$

However, due to the richness of the collection of atoms, this least squares estimator can easily overfit the data. To address this issue, one needs to control the complexity of the functions under consideration. In this framework, this can be achieved naturally by constraining the total variation of the signed measure ν in the representation (1.3). Specifically, we define the complexity of a function $f \in \mathcal{F}$ as

$$V(f) = \inf\big\{|\nu|(D) : f_\nu \equiv f\big\}. \tag{1.4}$$

Recall that for a finite signed measure ν on a measurable space Ω with Jordan decomposition $\nu = \nu^+ - \nu^-$, the variation of ν is defined as $|\nu| = \nu^+ + \nu^-$, and the total variation of ν is given by $|\nu|(\Omega)$. Here, the infimum is taken because a function may admit multiple representations of the form (1.3). If the representation is unique, the infimum can be dropped. This is the case for MARS, but not for XGBoost. One may also choose to leave certain atoms unconstrained and exclude them from the definition of the complexity measure. For MARS, we will leave the constant term and multi-affine terms unconstrained, while for XGBoost, only the constant term is excluded from the complexity measure (see Chapters 3 and 5).

This complexity measure $V(\cdot)$ can be viewed as a continuous extension, or an infinite-dimensional analogue, of the ℓ^1 norm on finite-dimensional vectors. For ease of discussion, let us momentarily assume that the representation (1.3) is unique, so that the infimum in (1.4) is not needed. In this case, if the signed measure ν is discrete and supported on $\mathbf{t}_1, \dots, \mathbf{t}_K$ with $\nu(\{\mathbf{t}_k\}) = c_k$ so that f_ν reduces to (1.2), the complexity of f_ν coincides with the ℓ^1 norm of the coefficient vector $(c_1, \dots, c_K)$:

$$V(f_\nu) = \sum_{k=1}^{K} |c_k|.$$

An analogous result continues to hold even when the representation is not unique and the infimum is necessary (see Theorem 14).

With this complexity measure, we define the constrained least squares estimator as

$$\hat{f}^d_{n,V} \in \operatorname*{argmin}_f \left\{ \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}^{(i)})\big)^2 : f \in \mathcal{F} \text{ and } V(f) \le V \right\}, \tag{1.5}$$

where $V > 0$ is a tuning parameter. For MARS, this estimator yields a variant of the original method that is more theoretically tractable, albeit computationally more expensive (see Chapter 3). For XGBoost, this estimator corresponds to the target optimization problem described in the original paper [29], which is approximately solved via their greedy boosting algorithm (see Chapter 5).

In light of the connection between the complexity measure $V(\cdot)$ and the ℓ^1 norm, the optimization problem (1.5) can be viewed as an instance of an infinite-dimensional lasso problem. Similar infinite-dimensional lasso frameworks have been extensively studied in the inverse problems literature, particularly in the context of super-resolution, where one seeks to recover high-resolution signals from low-resolution observations (see, e.g., [35, 18, 11, 24, 142, 45, 36, 37, 119, 33, 38, 118]). In this line of work, inverse problems over the space of measures are formulated as infinite-dimensional analogues of finite-dimensional ℓ^1 regularization problems, in a manner analogous to our setting. The total variation of measures is used as a regularizer extending the ℓ^1 norm to this setting, thereby promoting sparsity at the level of measures. Such problems, particularly in their regularized form, are also commonly referred to as Beurling lasso (BLASSO) in the literature. Infinite-dimensional lasso frameworks have also appeared in the statistical learning and approximation theory literature [125, 4, 111, 117, 114, 115, 116, 147, 135], though with different choices of atoms than in our work, most commonly those associated with single-hidden-layer neural networks.

Since the function class $\mathcal{F}$ is defined in a constructive manner, it is often not straightforward to determine whether a given function belongs to $\mathcal{F}$ or not. Moreover, the constructive definition provides limited insight into how similar or different $\mathcal{F}$ and the complexity $V(\cdot)$ are to other function classes and complexity measures studied in the literature. One of the main contributions of this thesis is to provide alternative characterizations of $\mathcal{F}$ and $V(\cdot)$ in terms of smoothness of functions.

We show that, for MARS and XGBoost, the complexity measure $V(\cdot)$, and hence the function class $\mathcal{F}$, can be characterized using mixed partial derivatives

$$f^{(\mathbf{p})} := \frac{\partial^{p_1+\cdots+p_d} f}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \tag{1.6}$$

with a fixed maximum order $\max_j p_j$. Specifically, for MARS, we show that the complexity of f can be characterized by the L^1 norms of mixed partial derivatives of f of maximum order two (see Section 3.1). For XGBoost, we show that the complexity measure is closely related to the L^1 norms of mixed partial derivatives of maximum order one (see Section 5.2). These results imply that, for a sufficiently smooth function f, it belongs to $\mathcal{F}$ if and only if it has bounded L^1 norms for mixed partial derivatives of maximum order one (for XGBoost) or two (for MARS). This suggests that the function class $\mathcal{F}$ becomes increasingly restrictive as the dimension d increases. Note, in particular, that the total order $\sum_j p_j$ of partial derivatives in the characterization can be as large as $2d$ for MARS and d for XGBoost.

However, these results should be interpreted with some caution. First, although the total order of partial derivatives scales with d, the order of each partial derivative is always bounded by one (for XGBoost) or two (for MARS), which prevents the function class $\mathcal{F}$ from becoming overly restrictive. More importantly, the above characterizations apply only to smooth functions. The function class $\mathcal{F}$ also contains non-smooth functions for which mixed partial derivatives are not even defined. Indeed, the atoms $A_{\mathbf{t}}$ and their finite linear combinations (1.2), which may constitute a particularly important subclass of $\mathcal{F}$, are non-differentiable in the case of MARS and discontinuous in the case of XGBoost.

Such L^q norm constraints on mixed partial derivatives, and the corresponding function classes, have been extensively studied in approximation theory (see, e.g., [40, 23, 43, 143]), where they are known to mitigate the curse of dimensionality to some extent from the perspective of approximation and interpolation. However, they have received comparatively little attention and remain largely underexplored in statistics. In our setting, the alternative characterizations of the function class $\mathcal{F}$ and the complexity measure $V(\cdot)$ via mixed partial derivatives reveal a previously underappreciated connection between MARS and XGBoost and nonparametric regression methods based on smoothness constraints. They provide a novel perspective on MARS and XGBoost, which are often regarded as purely algorithmic.

Instead of imposing the total variation constraint, one may impose a sign constraint on the measure ν. In other words, one may restrict ν in the representation (1.3) to be a nonnegative (or nonpositive) measure. In the discrete formulation (1.2), this corresponds to nonnegative (or nonpositive) finite linear combinations of atoms. Such constraints control the *shape* of the function rather than its complexity, thereby leading to shape-constrained regression methods. Shape-constrained regression seeks to estimate functions under qualitative constraints such as monotonicity and convexity, often without requiring tuning parameters (see, e.g., [67, 129] and references therein).

Applying this idea to the collection of atoms underlying MARS, we obtain a novel multivariate concave (or convex) regression method. This method is closely related to a less commonly used notion of concavity known as total concavity [120, 54], and is therefore referred to as totally concave regression. Total concavity is a multivariate generalization of univariate concavity. However, unlike classical notions of multivariate concavity, it is not characterized by the Hessian of functions. Instead, as in the case of MARS, it admits a characterization in terms of mixed partial derivatives of maximum order two. The difference from MARS is that these mixed partial derivatives are now required to be nonpositive, rather than having their L^1 norms constrained. Totally concave regression overcomes several limitations of existing multivariate concave (or convex) regression methods based on classical notions of concavity, and provides a compelling alternative in multi-dimensional settings. We introduce and study this method in detail in Chapter 4.

One challenge associated with the optimization problem (1.5), as well as its counterpart with a nonnegative measure (without the total variation constraint), is the computation of the estimator. These problems are infinite-dimensional, in the sense that the number of parameters to be determined is infinite. One may wish to invoke representer theorems for inverse problems on the space of measures, such as those in [148, 47, 15, 16], possibly under some additional assumptions on the parameter space D. However, unlike representer theorems for kernel methods in reproducing kernel Hilbert spaces, these results only guarantee the existence of sparse solutions of the form (1.2), and do not provide guidance on the optimal locations of $\mathbf{t}_1, \dots, \mathbf{t}_K$.

Fortunately, for MARS and totally concave regression, the solution to (1.5), as well as its sign-constrained variant, can be computed via convex optimization. More precisely, the measure ν in (1.3) can be restricted to be a discrete measure supported on a finite set with a lattice structure determined by the observations $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ (see Section 3.2). However, the resulting finite-dimensional convex optimization problems often involve a large number of parameters, and thus remain computationally demanding, especially when the number of observations n and the covariate dimension d are large. To address this issue, we also develop an approximate method for both MARS and totally concave regression, based on a coarse proxy lattice support for the measure. This approximation can also be computed via convex optimization and is scalable to larger datasets.

The same idea applies to XGBoost. However, in contrast to MARS and totally concave regression, where we develop new variants and methods based on (1.5), in the case of XGBoost, (1.5) corresponds precisely to the target optimization problem described in the original XGBoost paper. Since our focus is not on developing alternative computational methods for XGBoost, we do not pursue this direction further.

One of the main contributions of this thesis is a theoretical analysis of the accuracy of the estimator $\hat{f}^d_{n,V}$ (defined in (1.5)) for all three methods: MARS, XGBoost, and totally concave regression. In particular, under the standard regression model with either fixed or random design for $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, and under the well-specified assumption that the true regression function belongs to $\mathcal{F}$, we show that the estimator $\hat{f}^d_{n,V}$ achieves nearly dimension-free rates of convergence (see Sections 3.3, 4.3, and 5.3). Specifically, we prove that the rate of convergence of $\hat{f}^d_{n,V}$ is $n^{-4/5}$ for MARS and totally concave regression, and $n^{-2/3}$ for XGBoost, up to logarithmic factors whose exponents depend on d. These rates indicate that $\hat{f}^d_{n,V}$ avoids the curse of dimensionality to some extent.

These results also suggest that the function class $\mathcal{F}$ is sufficiently regularized in higher dimensions, which is consistent with the smoothness characterizations in terms of mixed partial derivatives (1.6). In fact, the above rates can be explained at an intuitive level through these characterizations. Recall that the total order of partial derivatives in the characterizations can be as large as $2d$ for MARS and d for XGBoost. Substituting these values into the classical nonparametric rate of convergence $n^{-2q/(2q+d)}$, with $q = 2d$ for MARS and $q = d$ for XGBoost, yields the above convergence rates (up to logarithmic factors).

Taken together, our results suggest that our approach and the corresponding estimator $\hat{f}^d_{n,V}$, particularly in the case of MARS and totally concave regression, can serve as a useful regression method. More broadly, they also provide theoretical insights, albeit indirect, into the empirical success of MARS and XGBoost in practice, which has thus far received relatively limited theoretical understanding.

The remainder of the thesis is organized as follows. In Chapter 2, we briefly review Hardy–Krause variation, which plays a key role in the alternative characterizations of the function classes $\mathcal{F}$ and the complexity measures $V(\cdot)$ for MARS and XGBoost developed in the following chapters. Chapter 3 presents results on MARS from [79], while Chapter 4 studies totally concave regression and presents results from [80]. Chapter 5 is devoted to our analysis of XGBoost from [81]. The thesis concludes with a discussion of the applicability and limitations of our approach.

Chapter 2

Hardy–Krause Variation

Before we study MARS and XGBoost through the infinite-dimensional lasso framework introduced in the previous chapter, we briefly review Hardy–Krause variation and some of its properties. Hardy–Krause variation is a multivariate generalization of univariate total variation, and it plays a key role in alternative characterizations of the function classes $\mathcal{F}$ (defined in (1.3)) and the complexity measures $V(\cdot)$ (defined in (1.4)) for MARS and XGBoost. Specifically, it connects the constructive definitions of $\mathcal{F}$ and $V(\cdot)$, given in terms of signed measures, with characterizations based on smoothness, particularly those involving mixed partial derivatives (1.6).

Hardy–Krause variation is typically defined for functions on bounded axis-aligned rectangles, such as $[0,1]^d$. We will need this standard version for MARS in Chapter 3, where the covariate domain will be restricted to $[0,1]^d$. Standard references for such Hardy–Krause variation include [93, 113, 1]. However, we will also need Hardy–Krause variation on the entire space $\mathbb{R}^d$ for XGBoost in Chapter 5, where no further restriction will be imposed on the covariate domain. We therefore first review the standard definition on compact domains in Section 2.1, and then introduce a natural extension to $\mathbb{R}^d$ in Section 2.2. We also present a key property of Hardy–Krause variation that enables it to play its role in the infinite-dimensional lasso framework.

2.1 On Bounded Axis-Aligned Rectangles

We first recall Vitali variation and several related concepts, which form a key building block for the definition of Hardy–Krause variation.

Definition 1 (Quasi-volume). *For $(u_1, \dots, u_m)$, $(v_1, \dots, v_m) \in \mathbb{R}^m$ with $u_j < v_j$ for all $j \in [m] := \{1, \dots, m\}$, the quasi-volume of a real-valued function g over the axis-aligned rectangle $\prod_{j=1}^m [u_j, v_j]$ is defined by*

$$\Delta\Big(g; \prod_{j=1}^m [u_j, v_j]\Big) = \sum_{\boldsymbol{\delta} \in \{0,1\}^m} (-1)^{\delta_1 + \dots + \delta_m} \cdot g\big((1-\delta_1)v_1 + \delta_1 u_1, \dots, (1-\delta_m)v_m + \delta_m u_m\big).$$

Definition 2 (Axis-aligned split)**.** *Let $(u_1, \ldots, u_m)$ and $(v_1, \ldots, v_m)$ be vectors in $\mathbb{R}^m$ with $u_j < v_j$ for all j. A collection $\mathcal{P}$ of subsets of $\prod_{j=1}^m [u_j, v_j]$ is called an axis-aligned split if it consists of rectangles of the form*

$$\prod_{j=1}^{m} \left[w_{l_j-1}^{(j)}, w_{l_j}^{(j)}\right] \quad \text{for } l_j \in [n_j] \text{ and } j \in [m], \tag{2.1}$$

where, for each $j \in [m]$, $u_j = w_0^{(j)} < w_1^{(j)} < \cdots < w_{n_j}^{(j)} = v_j$ is a partition of $[u_j, v_j]$.

Definition 3 (Vitali variation on bounded axis-aligned rectangles)**.** *The Vitali variation of a real-valued function g on $\prod_{j=1}^m [u_j, v_j]$ is defined by*

$$Vit\Big(g; \prod_{j=1}^{m} [u_j, v_j]\Big) = \sup_{\mathcal{P}} \sum_{R \in \mathcal{P}} |\Delta(g; R)|,$$

where the supremum is taken over all axis-aligned splits $\mathcal{P}$ of $\prod_{j=1}^m [u_j, v_j]$.

Vitali variation is closely related to mixed partial derivatives (1.6). Specifically, if g is sufficiently smooth, then the Vitali variation of g on $\prod_{j=1}^m [u_j, v_j]$ admits the following representation (see, e.g., [113, Section 9]):

$$\mathrm{Vit}\Big(g; \prod_{j=1}^{m} [u_j, v_j]\Big) = \int_{\prod_{j=1}^m [u_j, v_j]} \left| \frac{\partial^m g(\mathbf{x})}{\partial x_1 \cdots \partial x_m} \right| d\mathbf{x}. \tag{2.2}$$

We are now ready to define Hardy–Krause variation on bounded axis-aligned rectangles. Its definition requires specifying an anchor point among the vertices of the rectangle. For example, when the domain is $[0,1]^d$, the anchor point is chosen from $\{0,1\}^d$, and it is commonly taken to be the lower-left corner $\mathbf{0} = (0, \ldots, 0)$.

Let f be a real-valued function on $\prod_{j=1}^d [u_j, v_j]$, and let $\mathbf{a} = (a_1, \ldots, a_d) \in \prod_{j=1}^d \{u_j, v_j\}$ be the anchor point. For each $S \subseteq [d]$, define

$$f^S_{(a_j, j \in S^c)}(x_j, j \in S) = f(x_j, j \in S; a_j, j \in S^c) \quad \text{for } (x_j, j \in S) \in \prod_{j \in S} [u_j, v_j].$$

This function may be viewed as the restriction of f to the section of the domain obtained by fixing the coordinates in S^c at the anchoring values a_j.

Definition 4 (Hardy–Krause variation on bounded axis-aligned rectangles)**.** *The Hardy–Krause variation of a real-valued function f on $\prod_{j=1}^d [u_j, v_j]$ anchored at $\mathbf{a}$ is defined by*

$$HK_{\mathbf{a}}\Big(f; \prod_{j=1}^{d} [u_j, v_j]\Big) = \sum_{\emptyset \neq S \subseteq [d]} Vit\Big(f^S_{(a_j, j \in S^c)}; \prod_{j \in S} [u_j, v_j]\Big),$$

where the summation is taken over all nonempty subsets $S \subseteq [d]$.

In words, the Hardy–Krause variation of f is the sum of the Vitali variations of restrictions of f to sections of the domain obtained by anchoring some coordinates at a_j. This explains the terminology "anchor" for $\mathbf{a}$. For sufficiently smooth functions f, applying (2.2) to each term yields

$$\mathrm{HK}_{\mathbf{a}}\Big(f;\prod_{j=1}^{d}[u_j,v_j]\Big) = \sum_{\emptyset\neq S\subseteq[d]} \int_{\prod_{j\in S}[u_j,v_j]} \Big|\frac{\partial^{|S|}}{\prod_{j\in S}\partial x_j} f^{S}_{(a_j,j\in S^c)}(x_j, j\in S)\Big|\, d(x_j, j\in S). \quad (2.3)$$

Observe that the integrands in (2.3) are mixed partial derivatives of f in which each coordinate is differentiated at most once, while the coordinates that are not differentiated are anchored at a_j. This characterization of Hardy–Krause variation for smooth functions highlights the close connection between Hardy–Krause variation and mixed partial derivatives (1.6) of maximum order one.

This classical version of Hardy–Krause variation has also recently been used for nonparametric regression. In particular, [46] studied least squares estimation under a Hardy–Krause variation constraint, which they referred to as Hardy–Krause variation denoising. Hardy–Krause variation denoising can also be formulated within the infinite-dimensional lasso framework (1.5). The atoms of the method are functions on $[0,1]^d$ of the form

$$A_{\mathbf{t}}(\mathbf{x}) = \prod_{j=1}^{d} \mathbf{1}(x_j \geq t_j), \quad \text{for } \mathbf{t} = (t_1,\ldots,t_d) \in [0,1]^d,$$

and the estimator is a finite linear combination of these atoms (and hence piecewise constant). For functions represented as infinite linear combinations (1.3) of $A_{\mathbf{t}}$, the associated complexity measure (1.4) (excluding the constant term corresponding to $A_{\mathbf{0}}$) coincides with Hardy–Krause variation.

Moreover, in light of the smoothness characterization (2.3), Hardy–Krause variation denoising can also be viewed as a nonparametric regression method that constrains mixed partial derivatives of maximum order one. As in the case of MARS and XGBoost, briefly discussed in the Introduction, this suggests that the associated function class becomes increasingly restrictive as the dimension d grows, and helps explain the dimension-free rate $n^{-2/3}$ (up to logarithmic factors) of the estimator established in [46].

A slight variant of Hardy–Krause variation denoising, together with the corresponding regression method, has also been studied in [8, 90, 130], among many others. In this literature, the complexity measure and estimation method are referred to as sectional variation and highly adaptive lasso, respectively. The main difference between Hardy–Krause variation and sectional variation is that the former leaves the intercept unconstrained, whereas the latter constrains the intercept as well.

In Chapter 3, we will use Hardy–Krause variation on $[0,1]^d$ to illustrate alternative characterizations of the function class (1.3) and the complexity measure (1.4) for MARS. In particular, we will see that the complexity measure $V(\cdot)$ can be viewed as a second-order analogue of Hardy–Krause variation, and the corresponding estimator (1.5) as a second-order version of Hardy–Krause variation denoising.

We close this section by presenting a property of Hardy–Krause variation that helps explain why it naturally appears in the context of the infinite-dimensional lasso framework (1.5). The following theorem shows that every right-continuous[1] function with finite Hardy–Krause variation anchored at the lower-left corner can be represented as the cumulative distribution function of a finite signed measure on the same domain. This result also establishes a precise relationship between the Hardy–Krause variation of the function and the total variation of the corresponding signed measure. This connection between Hardy–Krause variation and signed measures underlies its natural appearance in the infinite-dimensional lasso framework, where functions are represented through the integral representation (1.3) over signed measures and the complexity measure (1.4) is defined in terms of the total variation of the representing measure. Analogous results to Theorem 1 also hold for other choices of the anchor point under slightly different one-sided continuity conditions.

Theorem 1 (Theorem 3 of [1])**.** *Suppose* $f : \prod_{j=1}^{m}[u_j, v_j] \to \mathbb{R}$ *is right-continuous and has finite Hardy–Krause variation anchored at* $\mathbf{a} = (u_1, \dots, u_m)$*. Then, there exists a unique finite signed (Borel) measure* ν *on* $\prod_{j=1}^{m}[u_j, v_j]$ *such that*

$$f(x_1, \dots, x_m) = \nu\Big(\prod_{j=1}^{m}[u_j, x_j]\Big) \quad \text{for } (x_1, \dots, x_m) \in \prod_{j=1}^{m}[u_j, v_j]. \tag{2.4}$$

Conversely, if f *is of the form* (2.4)*, then* f *has finite Hardy–Krause variation anchored at* $\mathbf{a}$ *and*

$$|\nu|\bigg(\prod_{j=1}^{m}[u_j, v_j]\bigg) = HK_{\mathbf{a}}\Big(f; \prod_{j=1}^{m}[u_j, v_j]\Big) + |f(\mathbf{a})|.$$

2.2 On the Entire Space $\mathbb{R}^d$

The standard definition of Hardy–Krause variation on bounded axis-aligned rectangles is sufficient for our study of MARS in Chapter 3. However, for our study of XGBoost in Chapter 5, where we work with functions defined on the entire space $\mathbb{R}^d$, we need a notion of Hardy–Krause variation on $\mathbb{R}^d$. We therefore make slight modifications to the standard definitions introduced in the previous section to accommodate the unbounded domain $\mathbb{R}^d$. We begin with the definition of Vitali variation.

Definition 5 (Vitali variation on $\mathbb{R}^m$)**.** *The Vitali variation of a real-valued function* g *on the entire space* $\mathbb{R}^m$ *is defined by*

$$Vit(g) = \sup_{\prod_{j=1}^{m}[u_j, v_j] \subseteq \mathbb{R}^m} Vit\Big(g; \prod_{j=1}^{m}[u_j, v_j]\Big).$$

[1]Throughout this thesis, the term “right(left)-continuous” refers to coordinate-wise right(left)-continuity.

Similarly to (2.2), for sufficiently smooth functions g, the Vitali variation of g on $\mathbb{R}^m$ can be expressed as

$$\mathrm{Vit}(g) = \int_{\mathbb{R}^m} \left| \frac{\partial^m g(\mathbf{x})}{\partial x_1 \cdots \partial x_m} \right| d\mathbf{x}. \tag{2.5}$$

Using this Vitali variation, we define Hardy–Krause variation on the entire space $\mathbb{R}^d$. As in the case of bounded axis-aligned rectangles, an anchor point must be specified. In the present setting, where there is no natural boundary, the anchor point needs to be placed at infinity (either $-\infty$ or $+\infty$), and this requires functions to be suitably well behaved at infinity, in the sense described below.

Let $\mathbf{a} = (a_1, \dots, a_d) \in \{-\infty, +\infty\}^d$ denote the anchor point. For each coordinate, there are two possible choices: $-\infty$ or $+\infty$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, a subset $S \subseteq [d]$ with $S^c := [d] \setminus S$, and $(x_j, j \in S) \in \mathbb{R}^{|S|}$, define

$$f^S_{(a_j, j \in S^c)}(x_j, j \in S) = \lim_{(x_j, j \in S^c) \to (a_j, j \in S^c)} f(x_1, \dots, x_d). \tag{2.6}$$

Rather than fixing the coordinates in S^c at the anchoring values a_j, here we take limits in the directions determined by a_j. For each $S \subseteq [d]$, we say that the function $f^S_{(a_j, j \in S^c)}$ is well defined if the above limit exists and is finite for all $(x_j, j \in S) \in \mathbb{R}^{|S|}$. When these functions are well defined for all $S \subseteq [d]$, Hardy–Krause variation on $\mathbb{R}^d$ is defined as follows.

Definition 6 (Hardy–Krause variation on $\mathbb{R}^d$)**.** *Fix $\mathbf{a} \in \{-\infty, +\infty\}^d$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function for which $f^S_{(a_j, j \in S^c)}$ is well defined for all $S \subseteq [d]$. The Hardy–Krause variation of f anchored at $\mathbf{a}$ is defined by*

$$HK_{\mathbf{a}}(f) = \sum_{\emptyset \neq S \subseteq [d]} Vit(f^S_{(a_j, j \in S^c)}).$$

As in (2.3), when f is sufficiently smooth, (2.5) implies that $\mathrm{HK}_{\mathbf{a}}(f)$ admits a representation as a sum of integrals of mixed partial derivatives of maximum order one:

$$\mathrm{HK}_{\mathbf{a}}(f) = \sum_{\emptyset \neq S \subseteq [d]} \int_{\mathbb{R}^{|S|}} \left| \frac{\partial^{|S|}}{\prod_{j \in S} \partial x_j} f^S_{(a_j, j \in S^c)}(x_j, j \in S) \right| d(x_j, j \in S). \tag{2.7}$$

In Chapter 5, we will see that the function class (1.3) and the complexity measure (1.4) for XGBoost are closely related to Hardy–Krause variation on $\mathbb{R}^d$. Specifically, we will show that the function class can be precisely characterized as the class of functions with finite Hardy–Krause variation satisfying certain one-sided continuity conditions, and that the complexity measure is within a constant factor (depending on d) of Hardy–Krause variation.

Chapter 3

Multivariate Adaptive Regression Splines (MARS)

Multivariate adaptive regression splines (MARS) [53] is a classical machine learning method for regression. Given data $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$, where $\mathbf{x}^{(i)} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, MARS fits a finite (often sparse) linear combination of atoms of the form

$$(x_1, \ldots, x_d) \mapsto \prod_{j \in S} a_{t_j}(x_j) \tag{3.1}$$

where $S \subseteq [d] := \{1, \ldots, d\}$ and

$$a_t(x) = (x - t)_+ \quad \text{or} \quad a_t(x) = (t - x)_+ \quad \text{for } x \in \mathbb{R}.$$

Here, $(\cdot)_+ := \max(\cdot, 0)$, and the function a_t is called a hinge function (often referred to as a ReLU function in the machine learning literature) with knot t. In words, the atoms of MARS consist of hinge functions along each coordinate and their products. Using these atoms, MARS aims to capture nonlinear and nonadditive relationships between the response y and the d-dimensional covariate $\mathbf{x} = (x_1, \ldots, x_d)$. The nonlinearity of the hinge functions and their product form enable MARS to model nonlinear patterns and interactions among covariates, while the simple structure of the atoms helps preserve model interpretability.

To fit a finite linear combination of these atoms, MARS employs a greedy forward-backward procedure similar in spirit to stepwise linear regression. Starting from a constant function, MARS iteratively augments the current model by adding new atoms chosen to achieve the largest reduction in residual sum of squares among a candidate set of atoms of the form (3.1). Typically, one also restricts attention to atoms involving at most s hinge functions for a pre-specified integer $s \geq 1$ (most commonly, $s = 2$), that is, the atoms (3.1) with $|S| \leq s$, where $|S|$ denotes the cardinality of S. In words, this corresponds to restricting attention to interactions among covariates of order at most s. After a sufficiently rich collection of atoms has been generated in this forward stage, MARS applies a backward pruning procedure to remove superfluous atoms from the model. We refer to [53, Section 9.4] for a detailed description of the MARS fitting algorithm.

Although MARS has enjoyed considerable success and has been widely used since its introduction in 1991, the greedy nature of its algorithm has made rigorous statistical analysis of the method challenging. In this chapter, building on the infinite-dimensional lasso framework via measures discussed in the Introduction, we propose a variant of MARS that can be computed through convex optimization and is substantially more amenable to theoretical analysis. We further show that this variant often performs comparably to, and sometimes better than, the usual MARS algorithm in practice, albeit at a higher computational cost. Through both theoretical analysis and numerical experiments, we demonstrate that our lasso-based variant of MARS can serve as an attractive alternative to the classical MARS algorithm, much as lasso serves as an alternative to stepwise linear regression.

In order to apply the lasso-based framework to MARS, we first normalize each covariate so that its domain becomes $[0, 1]$. Equivalently, we assume throughout that the regression function is defined on $[0, 1]^d$. In practice, this can be achieved by subtracting the smallest possible value from each covariate and dividing by its range. Such normalization places all covariates on a comparable scale, allowing the use of L^1-type coefficient constraints in a meaningful way. Without such normalization, the covariates $x_1, \dots, x_d$ may lie on very different scales, in which case constraining the L^1 norm of the coefficients of the atoms could be unnatural. After fitting a function on the transformed domain $[0, 1]^d$, one can recover the corresponding fitted function on the original scale by inverting the normalization.

Once we assume that each covariate x_j belongs to $[0, 1]$, we can make two simplifications to the collection of atoms.

- Instead of considering both types of hinge functions $(\cdot - t)_+$ and $(t - \cdot)_+$, we can restrict attention to $(\cdot - t)_+$. For $x \in [0, 1]$, we have

$$(t - x)_+ = (x - t)_+ - x + t = (x - t)_+ - (x - 0)_+ + t,$$

  and thus, every linear combination of atoms of the form (3.1) can be written as a linear combination of atoms of the same form in which a_t is always of the form $(\cdot - t)_+$.

- We can further assume that the knot t of a hinge function lies in $[0, 1)$. Indeed, if $t \geq 1$, then $(x - t)_+ = 0$ for all $x \in [0, 1]$, while if $t < 0$, then

$$(x - t)_+ = (x - 0)_+ - t,$$

  which is a linear combination of $(x - 0)_+$ and the constant function 1.

Moreover, following the classical algorithm of MARS, we restrict attention to atoms involving at most s hinge functions for a pre-specified integer $s \geq 1$. After these simplifications, the collection of MARS atoms reduces to

$$A_{(t_j, j \in S)}(x_1, \dots, x_d) = \prod_{j \in S} (x_j - t_j)_+ \quad \text{for } (x_1, \dots, x_d) \in [0, 1]^d, \tag{3.2}$$

where S is a subset of $[d]$ with $|S| \leq s$ and $t_j \in [0, 1)$ for each $j \in S$.

We now apply the lasso-based framework to this collection of atoms. Rather than using a single unified signed measure ν as in (1.3), we split it, for notational convenience, into multiple signed measures, each corresponding to a subcollection of atoms grouped by the set of involved covariates. More precisely, we partition the collection of the MARS atoms (3.2) according to the set S, and for each S, we use a signed measure ν_S to parametrize the corresponding subcollection. This yields functions of the form

$$f_{c,\{\nu_S\}}(x_1, \dots, x_d) := c + \sum_{0<|S|\le s} \int_{[0,1)^{|S|}} A_{(t_j, j\in S)}(x_1, \dots, x_d)\, d\nu_S(t_j, j \in S), \tag{3.3}$$

where $c \in \mathbb{R}$, the summation is over all nonempty subsets $S \subseteq [d]$ satisfying $|S| \le s$, and ν_S is a finite signed (Borel) measure on $[0,1)^{|S|}$. We denote the collection of all such functions by $\mathcal{F}^{d,s}_{\infty-\text{mars}}$. The subscript ∞ emphasizes that $\mathcal{F}^{d,s}_{\infty-\text{mars}}$ is defined via infinite-dimensional linear combinations of MARS atoms.

We measure the complexity of a function $f \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$ as in (1.4). We choose not to constrain the constant term and the multi-affine terms, which correspond to $(t_j, j \in S) = \mathbf{0} := (0, j \in S)$ for each S. Hence, we define the complexity measure on $\mathcal{F}^{d,s}_{\infty-\text{mars}}$ by

$$V_{\text{mars}}(f_{c,\{\nu_S\}}) = \sum_{0<|S|\le s} |\nu_S|\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big). \tag{3.4}$$

It can be shown that each function $f \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$ admits a unique representation of the form (3.3). This is why, unlike in (1.4), there is no need to take an infimum (over all possible representations) in the definition of (3.4).

Our variant of MARS based on the infinite-dimensional lasso framework (1.5) can then be formally written as

$$\hat{f}^{d,s}_{n,V} \in \operatorname*{argmin}_f \bigg\{ \sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)})\big)^2 : f \in \mathcal{F}^{d,s}_{\infty-\text{mars}} \text{ and } V_{\text{mars}}(f) \le V \bigg\}, \tag{3.5}$$

where $V > 0$ is a tuning parameter. The main goal of this chapter is to study this variant of MARS and show through both theoretical analysis and numerical experiments that it provides a theoretically tractable and practically competitive alternative to the classical MARS estimator.

In Section 3.1, we provide alternative characterizations of $\mathcal{F}^{d,s}_{\infty-\text{mars}}$, $V_{\text{mars}}(\cdot)$, and $\hat{f}^{d,s}_{n,V}$ in terms of smoothness, complementing the constructive definitions above. In particular, we show that they admit characterizations in terms of mixed partial derivatives (1.6) of maximum order two. Our alternative characterizations also reveal connections between our method and classical nonparametric regression methods based on smoothness constraints. Specifically, we show that our method can be viewed as a multivariate generalization of piecewise linear locally adaptive regression splines [100] and as a higher-order version of Hardy–Krause variation denoising [46].

In Section 3.2, we consider computational aspects of our estimator $\hat{f}^{d,s}_{n,V}$. We show that the optimization problem (3.5), which is formulated over a continuous object (a measure), can be reduced to a finite-dimensional convex optimization problem. Specifically, we show that it suffices to restrict attention to discrete signed measures supported on a lattice associated with the observations $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. We also discuss computational challenges involved in solving the resulting finite-dimensional convex optimization problem and introduce a computationally more efficient approximate version of our method.

In Sections 3.3 and 3.4, we establish the effectiveness of our method both theoretically and empirically. In Section 3.3, we study the theoretical accuracy of our estimator $\hat{f}^{d,s}_{n,V}$, along with its approximate version. In particular, we show that $\hat{f}^{d,s}_{n,V}$ achieves the dimension-free rate $n^{-4/5}$, up to logarithmic factors, under standard regression models in both fixed-design and random-design settings. We also show that this rate is nearly minimax optimal, again up to logarithmic factors. In Section 3.4, we conduct numerical experiments comparing $\hat{f}^{d,s}_{n,V}$ and its approximate version with the usual MARS estimator. We evaluate these methods on several simulated and real datasets and demonstrate that our methods often compare favorably with the usual MARS estimator.

The results of this chapter are drawn from [79]. We refer the reader to Appendix C for the proofs of all results presented in this chapter.

3.1 Characterizations in terms of Smoothness

In this section, we provide alternative characterizations of $\mathcal{F}^{d,s}_{\infty-\text{mars}}$, $V_{\text{mars}}(\cdot)$, and $\hat{f}^{d,s}_{n,V}$ in terms of smoothness. To motivate the results for general d and s, let us first consider the univariate case $d = s = 1$.

Smoothness Characterization for $d = s = 1$

For $d = s = 1$, $\mathcal{F}^{1,1}_{\infty-\text{mars}}$ consists of all functions $f : [0,1] \to \mathbb{R}$ of the form

$$f(x) = c + \int_{[0,1)} (x - t)_+ \, d\nu(t) \tag{3.6}$$

where c is a real number and ν is a finite signed measure on $[0,1)$, and the complexity measure of f is given by the variation of ν on $(0,1)$; that is, $V_{\text{mars}}(f) = |\nu|((0,1))$.

The following simple arguments show that $\mathcal{F}^{1,1}_{\infty-\text{mars}}$ can be characterized in terms of smoothness. First, by replacing $(x-t)_+$ with $\int_0^1 \mathbf{1}(t \le s \le x) \, ds$ in the integral in (3.6) and changing the order of integration, we obtain

$$f(x) = c + \int_0^x g(t) \, dt, \tag{3.7}$$

where the function $g : [0,1] \to \mathbb{R}$ is given by

$$g(t) = \nu([0,t] \cap [0,1)). \tag{3.8}$$

It can be readily verified that the function g in (3.8) is right-continuous on $[0,1]$ and left-continuous at 1, and the total variation $\mathrm{TV}(g)$ of g is finite and can be represented as

$$\mathrm{TV}(g) = |\nu|((0,1)).$$

Here, the total variation of a function $h : [0,1] \to \mathbb{R}$ is defined by

$$\mathrm{TV}(h) := \mathrm{TV}(h;[0,1]) = \sup_{0=z_0<z_1<\cdots<z_m=1} \sum_{k=1}^{m} \big|h(z_k) - h(z_{k-1})\big|,$$

where the supremum is over all integers $m \geq 1$ and partitions $0 = z_0 < z_1 < \cdots < z_m = 1$ of $[0,1]$. Conversely, every function $g : [0,1] \to \mathbb{R}$ that is right-continuous on $[0,1]$, left-continuous at 1, and has finite total variation can be written as (3.8) for a unique signed measure ν on $[0,1)$ (see, e.g., [1, Theorem 3]). Putting these observations together, we can argue that $\mathcal{F}^{1,1}_{\infty-\mathrm{mars}}$ has the following alternative characterization:

$$\begin{aligned} \mathcal{F}^{1,1}_{\infty-\mathrm{mars}} = \Big\{ f : [0,1] \to \mathbb{R} : \exists c \in \mathbb{R} \text{ and } g : [0,1] \to \mathbb{R} \text{ s.t.} \\ g \text{ is right-continuous on } [0,1], \text{ left-continuous at } 1, \\ \mathrm{TV}(g) < \infty, \text{ and } f(x) = c + \int_0^x g(t)\,dt \text{ for all } x \in [0,1] \Big\}. \end{aligned} \tag{3.9}$$

Moreover, we can see that the complexity measure $V_{\mathrm{mars}}(f)$ for $f \in \mathcal{F}^{1,1}_{\infty-\mathrm{mars}}$ is equal to the total variation $\mathrm{TV}(g)$ of the function g appearing in (3.9).

For every function $f \in \mathcal{F}^{1,1}_{\infty-\mathrm{mars}}$, we can show that g satisfying the conditions in (3.9) is unique, and thus, we can consider such g as a particular derivative of f satisfying (3.7). If we denote it by Df, the estimator $\hat{f}^{1,1}_{n,V}$ then can be alternatively written as

$$\hat{f}^{1,1}_{n,V} \in \operatorname*{argmin}_f \Big\{ \sum_{i=1}^{n} \big(y_i - f(x^{(i)})\big)^2 : \mathrm{TV}(Df) \leq V \Big\}.$$

The representation (3.7) implies, by the Lebesgue differentiation theorem (see, e.g., [126, Theorem 7.10]), that f' exists and is equal to Df almost everywhere (with respect to the Lebesgue measure) on $[0,1]$. Hence, we can also describe $\hat{f}^{1,1}_{n,V}$ somewhat loosely as

$$\hat{f}^{1,1}_{n,V} \in \operatorname*{argmin}_f \Big\{ \sum_{i=1}^{n} \big(y_i - f(x^{(i)})\big)^2 : \mathrm{TV}(f') \leq V \Big\}.$$

The corresponding penalized version

$$\operatorname*{argmin}_f \Big\{ \sum_{i=1}^{n} \big(y_i - f(x^{(i)})\big)^2 + \lambda \mathrm{TV}(f') \Big\} \tag{3.10}$$

was proposed by [100] as part of the class of estimators collectively called locally adaptive regression splines. In the univariate case, $\hat{f}^{1,1}_{n,V}$ can thus be seen as a constrained analogue of the locally adaptive regression spline estimator of [100] when the order k (in their notation) equals 2. [137] used the terminology *second-order total variation regularization*, and [82, 145] used the terminology *first-order trend filtering* for (3.10). Therefore, our estimator $\hat{f}^{d,s}_{n,V}$ can be viewed as a multivariate generalization of piecewise linear (second-order) locally adaptive regression splines, second-order total variation regularization, or first-order trend filtering.

There are other methods generalizing piecewise linear locally adaptive regression splines and first-order trend filtering to multivariate settings. See, for example, [114, 115] for some recent work on multivariate extensions of piecewise linear locally adaptive regression splines, and [112, 127] for recent progress on multivariate extensions of first-order trend filtering.

From the alternative characterization of $\mathcal{F}^{1,1}_{\infty-\mathrm{mars}}$ given above, it follows that every sufficiently smooth function $f : [0,1] \to \mathbb{R}$ belongs to $\mathcal{F}^{1,1}_{\infty-\mathrm{mars}}$. Indeed, if f' and f'' exist everywhere and are continuous on $[0,1]$, then we have

$$f(x) = f(0) + \int_0^x f'(t)\, dt$$

for all $x \in [0,1]$, and

$$\mathrm{TV}(f') = \int_0^1 |f''| < \infty.$$

Thus, in this case, f belongs to $\mathcal{F}^{1,1}_{\infty-\mathrm{mars}}$ and the complexity measure of f can be written as

$$V_{\mathrm{mars}}(f) = \int_0^1 |f''|. \tag{3.11}$$

This formula highlights the role of the second derivative f'' in the determination of $V_{\mathrm{mars}}(f)$ for each sufficiently smooth function f.

Smoothness Characterization for General d and s

As we have seen in the previous subsection, in the univariate case, $\mathcal{F}^{1,1}_{\infty-\mathrm{mars}}$ consists of all functions f satisfying (3.7) with some function g having finite total variation and some one-sided continuity. An analogous characterization holds for general d and s. For general d and s, the role of total variation in the univariate case is played by Hardy–Krause variation. More precisely, it is taken over by Hardy–Krause variation anchored at $\mathbf{0}$, which we denote by $\mathrm{HK}_{\mathbf{0}}(\cdot)$ and refer to as $\mathrm{HK}_{\mathbf{0}}$ variation. Recall the definition of Hardy–Krause variation and its properties from Section 2.1.

The following result provides an alternative characterization of $\mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ and $V_{\mathrm{mars}}(\cdot)$ in terms of smoothness.

Proposition 1. *The function class $\mathcal{F}^{d,s}_{\infty-mars}$ consists precisely of all functions of the form*

$$f(x_1, \dots, x_d) = c + \sum_{0<|S|\le s} \int_{\prod_{j\in S}[0,x_j]} g_S(t_j, j \in S)\, d(t_j, j \in S) \tag{3.12}$$

for some $c \in \mathbb{R}$ and some functions $\{g_S : S \subseteq [d], 0 < |S| \le s\}$, where for each S,

- *g_S is a real-valued function on $[0,1]^{|S|}$;*
- *$HK_{\mathbf{0}}(g_S) < \infty$;*
- *g_S is right-continuous on $[0,1]^{|S|}$;*
- *g_S is left-continuous at each point $(x_j, j \in S) \in [0,1]^{|S|} \setminus [0,1)^{|S|}$ with respect to all the j^{th} coordinates where $x_j = 1$.*

Furthermore, the complexity of f in (3.12) can be written in terms of the Hardy–Krause variations of g_S as

$$V_{mars}(f) = \sum_{0<|S|\le s} HK_{\mathbf{0}}(g_S).$$

Proposition 1 is completely analogous to (3.9) for the case $d = s = 1$. Specifically, condition (3.12) is analogous to condition (3.7). The condition $\mathrm{HK}_{\mathbf{0}}(g_S) < \infty$ for each S corresponds to the condition $\mathrm{TV}(g) < \infty$. The right-continuity of each g_S on $[0,1]^{|S|}$ is matched with the right-continuity on $[0,1]$. Lastly, the left-continuity of each g_S at each $(x_j, j \in S) \in [0,1]^{|S|} \setminus [0,1)^{|S|}$ (with respect to all the j^{th} coordinates where $x_j = 1$) is a counterpart of the left-continuity at 1. It is also interesting to note that $V_{\text{mars}}(f)$ equals the *sum* of the Hardy–Krause variations of g_S over all $S \subseteq [d]$ with $0 < |S| \le s$.

For every function $f \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$, it can be easily checked that g_S appearing in Proposition 1 are uniquely determined by f. As in the case $d = s = 1$, we can thus consider such g_S as particular derivatives of f satisfying (3.12). Let us denote them by $D^S f$ for each nonempty $S \subseteq [d]$ with $|S| \le s$. We can then write our estimator $\hat{f}^{d,s}_{n,V}$ alternatively as

$$\hat{f}^{d,s}_{n,V} \in \operatorname*{argmin}_f \bigg\{ \sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)})\big)^2 : \sum_{0<|S|\le s} \mathrm{HK}_{\mathbf{0}}(D^S f) \le V \bigg\}.$$

Hence, our estimator can be viewed as a least squares estimator under a specific smoothness constraint involving the sum of the Hardy–Krause variations of the particular derivatives defined via D^S.

Recall that the univariate condition (3.7) implies that f' exists and equals Df almost everywhere. Similarly, condition (3.12) imposes a certain kind of smoothness on f and characterizes the corresponding derivatives in terms of $D^S f$. For each nonempty $S \subseteq [d]$, let

$$D_S f(x_j, j \in S) = \lim_{\epsilon \to 0} \frac{1}{\epsilon^{|S|}} \cdot \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j\in S} \delta_j} \cdot f\big(\mathbf{x}(\widetilde{\mathbf{x} + \epsilon \mathbf{1}})_{\boldsymbol{\delta}}\big),$$

if the limit exists, where

$$\left(\mathbf{x}\widetilde{(\mathbf{x}+\epsilon\mathbf{1})}_{\boldsymbol{\delta}}\right)_j = \begin{cases} \delta_j x_j + (1-\delta_j)(x_j+\epsilon) & \text{if } j \in S \\ 0 & \text{otherwise} \end{cases}$$

for $j \in [d]$. For example, if $d = 3$ and $S = \{1, 2\}$, $D_{12}f$ is defined as

$$D_{12}f(x_1, x_2) = \lim_{\epsilon \to 0} \frac{1}{\epsilon^2}\big(f(x_1+\epsilon, x_2+\epsilon, 0) - f(x_1, x_2+\epsilon, 0) - f(x_1+\epsilon, x_2, 0) + f(x_1, x_2, 0)\big),$$

if the limit exists. In contrast to mixed partial derivative $f^{(\mathbf{p})}$ (defined in (1.6)), in which partial derivatives $\partial/\partial x_j$ are taken sequentially, here all the coordinates $j \in S$ are considered simultaneously. Also, note that the remaining coordinates $j \notin S$ are set to zero for $D_S f$.

As in the case $d = s = 1$, we can show that $D_S f$ exist and equal $D^S f$ almost everywhere (with respect to the Lebesgue measure) on $[0,1]^{|S|}$. The precise statement is given below.

Proposition 2. *Suppose that condition (3.12) holds. Then, for each nonempty subset $S \subseteq [d]$, $D_S f = 0$ if $|S| > s$, and $D_S f = D^S f$ almost everywhere (with respect to the Lebesgue measure) on $[0,1]^{|S|}$ if $|S| \le s$.*

Proposition 1 also implies that every sufficiently smooth function belongs to $\mathcal{F}^{d,d}_{\infty-\mathrm{mars}}$. This is proved in the next result, which also gives an expression for $V_{\mathrm{mars}}(f)$ in terms of the L^1 norms of the mixed partial derivatives of f, for sufficiently smooth functions f. The following notation will be used below. For each nonempty subset $S \subseteq [d]$, we let J_S be the set of all $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1,2\}^d$ such that

$$\max_j p_j = 2 \quad \text{and} \quad p_j = \begin{cases} 1 \text{ or } 2 & \text{if } j \in S \\ 0 & \text{if } j \notin S. \end{cases}$$

Also, for each $\mathbf{p} = (p_1, \dots, p_d) \in \mathbb{Z}^d_{\ge 0}$, we let

$$T^{(\mathbf{p})} = T^{(\mathbf{p})}_1 \times \cdots \times T^{(\mathbf{p})}_d \quad \text{where} \quad T^{(\mathbf{p})}_j = \begin{cases} [0,1] & \text{if } p_j = \max_k p_k \\ \{0\} & \text{otherwise.} \end{cases}$$

Lemma 1. *Suppose $f : [0,1]^d \to \mathbb{R}$ is smooth in the sense that $f^{(\mathbf{p})}$ exists and is continuous for every $\mathbf{p} \in \{0,1,2\}^d$. Then, $f \in \mathcal{F}^{d,d}_{\infty-mars}$ and*

$$V_{mars}(f) = \sum_{\substack{\mathbf{p} \in \{0,1,2\}^d \\ \max_j p_j = 2}} \int_{T^{(\mathbf{p})}} |f^{(\mathbf{p})}|. \tag{3.13}$$

Moreover, if $f^{(\mathbf{p})} = 0$ for all $\mathbf{p} \in \{0,1\}^d$ with $\sum_j p_j > s$ in addition, then $f \in \mathcal{F}^{d,s}_{\infty-mars}$ and

$$V_{mars}(f) = \sum_{0 < |S| \le s} \sum_{\mathbf{p} \in J_S} \int_{T^{(\mathbf{p})}} |f^{(\mathbf{p})}|. \tag{3.14}$$

Note that the integrals on the right-hand side of (3.13) and (3.14) are taken only over those coordinates j for which $p_j = 2$; the remaining coordinates are fixed at zero. In fact, Lemma 1 remains valid under weaker smoothness assumptions. However, since the precise degree of smoothness required is of little importance here, we impose conditions that are more than sufficient.

The formula (3.13) is a multivariate generalization of the univariate formula (3.11), stating that for sufficiently smooth functions f, $V_{\mathrm{mars}}(f)$ is the sum of the L^1 norms of the mixed partial derivatives of f of order 2, where we take at most two partial derivatives along each coordinate and exactly two partial derivatives along at least one coordinate. Note that mixed partial derivatives of total order $(p_1 + \cdots + p_d)$ up to $2d$ appear in (3.13). From this perspective, we can think of the smoothness order of our complexity measure $V_{\mathrm{mars}}(\cdot)$ as $2d$, which is proportional to the dimension d. This gives an intuitive explanation on why our estimators achieve the dimension-free rate $n^{-4/5}$ (up to logarithmic factors), as we will see in Section 3.3. However, we should also note that the maximum total order $2d$ is attained only by the mixed partial derivative $f^{(2,\dots,2)}$, since the order of each partial derivative is always capped at two, and this prevents our function class from being too small and restrictive. For these reasons, we can say that our complexity measure is an effective constraint that leads to estimators avoiding the curse of dimensionality (to some extent) while keeping the corresponding function class reasonably large.

There is also an interesting connection between $V_{\mathrm{mars}}(\cdot)$ and $\mathrm{HK}_{\mathbf{0}}(\cdot)$ via (3.13). Specifically, if the condition $\max_j p_j = 2$ in (3.13) is replaced with $\max_j p_j = 1$, then one obtains the formula for $\mathrm{HK}_{\mathbf{0}}(f)$ for sufficiently smooth functions f (recall (2.3)). Hence, we can also view $V_{\mathrm{mars}}(\cdot)$ as *second-order* Hardy–Krause variation (anchored at $\mathbf{0}$) and our method as a second-order analogue of Hardy–Krause variation denoising [46].

In Appendix A.1, we describe the results of this section by specializing to the case $d = s = 2$. We encourage the reader who finds the results in this section too abstract to refer to Appendix A.1 for more explicit formulae.

3.2 Existence, Computation, and Approximation

In this section, we prove the existence of our estimator $\hat{f}^{d,s}_{n,V}$ (defined in (3.5)) and show that it can be computed via convex optimization algorithms. We also introduce a computationally more efficient approximate version of our estimator.

We start with the observation that the objective function of the optimization problem defined in (3.5) only depends on the function f through its values at the design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. As proved in the next lemma, this observation allows us to restrict our attention to the finite-dimensional subclass of $\mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ consisting of the functions (3.3) where each ν_S is a discrete signed measure supported on the lattice generated by the design points. For each $j \in [d]$, let $\mathcal{V}_j$ denote the finite subset of $[0, 1]$ consisting of the points $0, x^{(1)}_j, \dots, x^{(n)}_j, 1$ (recall here that $x^{(i)}_j$ denotes the j^{th} coordinate of the i^{th} design point $\mathbf{x}^{(i)} = (x^{(i)}_1, \dots, x^{(i)}_d)$).

As there could be ties among $0, x_j^{(1)}, \ldots, x_j^{(n)}, 1$, we will write, for some $n_j \in \{1, \ldots, n+1\}$,

$$\mathcal{V}_j = \big\{v_0^{(j)}, v_1^{(j)}, \ldots, v_{n_j}^{(j)}\big\}, \quad \text{where } 0 = v_0^{(j)} < \cdots < v_{n_j}^{(j)} = 1.$$

Note, in particular, that the cardinality of $\mathcal{V}_j$ is $n_j + 1$, that $v_0^{(j)}$ is always 0, and that $v_{n_j}^{(j)}$ is always 1. The next lemma implies that, for the optimization problem (3.5), we can restrict to the functions of the form (3.3) where each ν_S is a discrete signed measure supported on the finite set $(\prod_{j\in S} \mathcal{V}_j) \cap [0,1)^{|S|}$.

Lemma 2. *Suppose we are given a real number c and a collection of finite signed measures $\{\nu_S\}$ where ν_S is defined on $[0,1)^{|S|}$ for each nonempty subset $S \subseteq [d]$ with $|S| \leq s$. Then, there exists a collection of discrete signed measures $\{\mu_S\}$ where μ_S is concentrated on $(\prod_{j\in S} \mathcal{V}_j) \cap [0,1)^{|S|}$ for each S such that*

- $f_{c,\{\mu_S\}}(\mathbf{x}^{(i)}) = f_{c,\{\nu_S\}}(\mathbf{x}^{(i)})$ *for all $i = 1, \ldots, n$;*
- $V_{mars}(f_{c,\{\mu_S\}}) \leq V_{mars}(f_{c,\{\nu_S\}})$.

When ν_S is concentrated on $(\prod_{j\in S} \mathcal{V}_j) \cap [0,1)^{|S|}$ for each S, the function $f_{c,\{\nu_S\}}$ can be written as

$$c + \sum_{0<|S|\leq s} \sum_{\mathbf{l}\in \prod_{j\in S}\{0,\ldots,n_j-1\}} \nu_S\big(\big\{\big(v_{l_j}^{(j)}, j\in S\big)\big\}\big) \cdot \prod_{j\in S}(x_j - v_{l_j}^{(j)})_+, \tag{3.15}$$

Also, its complexity measure becomes

$$V_{\text{mars}}(f_{c,\{\nu_S\}}) = \sum_{0<|S|\leq s} \sum_{\mathbf{l}\in \prod_{j\in S}\{0,\ldots,n_j-1\}\setminus\{\mathbf{0}\}} \big|\nu_S\big(\big\{\big(v_{l_j}^{(j)}, j\in S\big)\big\}\big)\big|.$$

The above function (3.15) is a linear combination of the atoms (3.1) whose knots t_j are chosen from $\mathcal{V}_j \setminus \{1\}$, and its complexity measure equals the absolute sum of the coefficients of the involved atoms with at least one nonlinear term. Thus, if we additionally assume that f in the problem (3.5) is constructed from discrete signed measures as above, then (3.5) reduces to a finite-dimensional lasso problem. Lemma 2 then implies that every solution to this finite-dimensional lasso problem is also a solution to (3.5). A precise statement is given in the following result.

Proposition 3. *Let*

$$J = \left\{(S, \mathbf{l}) : S \subseteq [d], 0 < |S| \leq s, \text{ and } \mathbf{l} \in \prod_{j\in S}\{0, \ldots, n_j - 1\}\right\}$$

and let $\mathbf{M}$ be the $n \times |J|$ matrix with columns indexed by $(S, \mathbf{l}) \in J$ such that

$$M_{i,(S,\mathbf{l})} = \prod_{j\in S}(x_j^{(i)} - v_{l_j}^{(j)})_+ \quad \text{for } i = 1, \ldots, n \text{ and } (S, \mathbf{l}) \in J.$$

Also, let $(\hat{c}, \hat{\boldsymbol{\gamma}}^{d,s}_{n,V}) \in \mathbb{R} \times \mathbb{R}^{|J|}$ *be a solution to the following finite-dimensional lasso problem*

$$(\hat{c}, \hat{\boldsymbol{\gamma}}^{d,s}_{n,V}) \in \underset{c\in\mathbb{R}, \boldsymbol{\gamma}\in\mathbb{R}^{|J|}}{\operatorname{argmin}} \left\{ \|y - c\mathbf{1} - \mathbf{M}\boldsymbol{\gamma}\|_2^2 : \sum_{\substack{(S,\mathbf{l})\in J \\ \mathbf{l}\neq\mathbf{0}}} |\gamma_{S,\mathbf{l}}| \le V \right\}, \tag{3.16}$$

where $\mathbf{1} := (1, \dots, 1)$ *and* $\mathbf{y} = (y_1, \dots, y_n)$ *is the vector of observations. Then, the function* f *on* $[0,1]^d$ *defined by*

$$f(x_1, \dots, x_d) = \hat{c} + \sum_{(S,\mathbf{l})\in J} (\hat{\gamma}^{d,s}_{n,V})_{S,\mathbf{l}} \cdot \prod_{j\in S} (x_j - v^{(j)}_{l_j})_+ \tag{3.17}$$

is a solution to the problem (3.5)*. The problem* (3.5) *can have multiple solutions, but every solution* $\hat{f}^{d,s}_{n,V}$ *satisfies*

$$\hat{f}^{d,s}_{n,V}(\mathbf{x}^{(i)}) = \hat{c} + (\mathbf{M}\hat{\boldsymbol{\gamma}}^{d,s}_{n,V})_i = \hat{c} + \sum_{(S,\mathbf{l})\in J} (\hat{\gamma}^{d,s}_{n,V})_{S,\mathbf{l}} \cdot \prod_{j\in S} (x^{(i)}_j - v^{(j)}_{l_j})_+$$

for every $i = 1, \dots, n$*.*

Because the set

$$\left\{ c\mathbf{1} + \mathbf{M}\boldsymbol{\gamma} : c \in \mathbb{R}, \boldsymbol{\gamma} \in \mathbb{R}^{|J|}, \text{ and } \sum_{\substack{(S,\mathbf{l})\in J \\ \mathbf{l}\neq\mathbf{0}}} |\gamma_{S,\mathbf{l}}| \le V \right\}$$

is closed and convex, there exists a solution to the finite-dimensional lasso problem (3.16). Hence, the existence of solutions to our estimation problem (3.5) is guaranteed by Proposition 3. Also, once we find a solution to the problem (3.16) via any convex optimization algorithms, we can construct a solution $\hat{f}^{d,s}_{n,V}$ to the problem (3.5) through equation (3.17).

However, solving the finite-dimensional lasso problem (3.16) can be computationally intensive if n is large because the number of columns of $\mathbf{M}$ equals

$$|J| = \sum_{0<|S|\le s} \prod_{j\in S} n_j,$$

which is of order $O(n^s)$ (ignoring a multiplicative factor in d) in the worst case when each $n_j = O(n)$. In the current implementation of our method, we utilize the convex optimization software MOSEK as a black-box tool for solving the problem (3.16) (see Section 3.4 for more details). Using this black-box tool involves creating the whole matrix $\mathbf{M}$, and thus, when n is large, our current implementation not only requires a large amount of space for this matrix but also often consumes most of running time constructing it.

This limitation motivates us to come up with the following approximate method. As we have seen above, Lemma 2 ensures that we only need to consider discrete signed measures ν_S supported on the lattices $(\prod_{j\in S} \mathcal{V}_j) \cap [0,1)^{|S|}$ for our estimation problem (3.5). In the approximate method, we instead restrict our attention to discrete signed measures ν_S supported on the lattices generated by

$$\widetilde{\mathcal{V}}_j = \left\{0, \frac{1}{N_j}, \frac{2}{N_j}, \ldots, 1\right\} \tag{3.18}$$

for some pre-selected positive integers $N_1, \ldots, N_d$, and we only take into consideration the atoms corresponding to those signed measures. Note that in contrast to $\mathcal{V}_j$ whose cardinality is of order $O(n)$ in the worst case, the cardinality of each set $\widetilde{\mathcal{V}}_j$ is always $N_j + 1$ regardless of the design points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$.

We then consider the finite-dimensional optimization problem to which the problem (3.5) reduces when we additionally impose such restrictions on signed measures ν_S. We call this problem the approximate (finite-dimensional optimization) problem. The approximate problem has the same form as (3.16) but with different $\mathbf{M}$ and J. Here,

$$J = \left\{(S, \mathbf{l}) : S \subseteq [d], 0 < |S| \leq s, \text{ and } \mathbf{l} \in \prod_{j \in S} \{0, \ldots, N_j - 1\}\right\}$$

and $\mathbf{M}$ is the $n \times |J|$ matrix with columns indexed by $(S, \mathbf{l}) \in J$ such that

$$M_{i,(S,\mathbf{l})} = \prod_{j \in S} \left(x_j^{(i)} - \frac{l_j}{N_j}\right)_+ \quad \text{for } i = 1, \ldots, n \text{ and } (S, \mathbf{l}) \in J.$$

As opposed to the original finite-dimensional problem (3.16), the number of columns of $\mathbf{M}$ in this problem is always fixed and not affected by the design points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$. Hence, this approximate problem can be solved much more efficiently than (3.16), especially when n is large. Once we find a solution to the approximate problem, we can construct an estimator of the true underlying function f^* through equation (3.17) as before. We denote this estimator by $\widetilde{f}^{d,s}_{n,V}$ and call it an approximate version of $\hat{f}^{d,s}_{n,V}$. In the next section, we study the theoretical accuracy of $\widetilde{f}^{d,s}_{n,V}$ along with $\hat{f}^{d,s}_{n,V}$. We will see that if we choose N_j appropriately, the approximate method is as accurate as the original method, while it significantly improves computational efficiency.

3.3 Theoretical Accuracy

This section is dedicated to the study of the theoretical accuracy of $\hat{f}^{d,s}_{n,V}$ and $\widetilde{f}^{d,s}_{n,V}$ as an estimator for unknown regression functions. We consider both fixed-design and random-design settings.

Fixed Design

Here, we assume that $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ form a lattice:

$$\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\} = \prod_{j=1}^{d} \left\{v^{(j)}_{l_j} : l_j \in \{0, \ldots, n_j - 1\}\right\} \tag{3.19}$$

where for every $j \in [d]$, we have $n_j \geq 2$, $0 = v_0^{(j)} < v_1^{(j)} < \cdots < v_{n_j-1}^{(j)} \leq 1$, and

$$v_{l_j}^{(j)} - v_{l_j-1}^{(j)} \geq \frac{\rho}{n_j} \quad \text{for all } l_j = 1, \ldots, n_j - 1$$

for some constant $\rho > 0$. Although this lattice design assumption is admittedly restrictive, it is fairly standard in the nonparametric function estimation literature (see, e.g., [109]). We also assume that $y_1, \ldots, y_n$ are generated according to the regression model

$$y_i = f^*(\mathbf{x}^{(i)}) + \xi_i$$

where $f^* : [0,1]^d \to \mathbb{R}$ is the unknown regression function and ξ_i are independent sub-Gaussian errors with mean zero and with a sub-Gaussian parameter σ, i.e.,

$$\mathbb{E}[e^{\lambda \xi_i}] \leq e^{\sigma^2 \lambda^2 / 2} \quad \text{for all } \lambda \in \mathbb{R}. \tag{3.20}$$

We measure the accuracy of an estimator $\hat{f}_n$ of f^* via the squared empirical L^2 norm

$$\|\hat{f}_n - f^*\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \big(\hat{f}_n(\mathbf{x}^{(i)}) - f^*(\mathbf{x}^{(i)})\big)^2 \tag{3.21}$$

and define its risk as

$$R_F(\hat{f}_n, f^*) = \mathbb{E}\|\hat{f}_n - f^*\|_n^2$$

where the expectation is taken over $y_1, \ldots, y_n$.

Our first result states an upper bound on the risk of $\hat{f}_{n,V}^{d,s}$ under the assumption $f^* \in \mathcal{F}_{\infty-\text{mars}}^{d,s}$ and $V_{\text{mars}}(f^*) \leq V$.

Theorem 2. *Suppose $f^* \in \mathcal{F}_{\infty-mars}^{d,s}$ and $V_{mars}(f^*) \leq V$. Also, assume the lattice design* (3.19) *and sub-Gaussian errors* (3.20)*. The estimator $\hat{f}_{n,V}^{d,s}$ then satisfies*

$$R_F(\hat{f}_{n,V}^{d,s}, f^*) \leq C_{\rho,d} \Big(\frac{\sigma^2 V^{1/2}}{n}\Big)^{4/5} \Big[\log\Big(2 + \frac{V n^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/5} + C_{\rho,d} \frac{\sigma^2}{n} (\log n)^2 \tag{3.22}$$

for some positive constant $C_{\rho,d}$ depending on ρ and d.

Note that for fixed ρ, d, σ, and V and sufficiently large n, the first term is the dominant term on the right-hand side of (3.22), so that

$$R_F(\hat{f}_{n,V}^{d,s}, f^*) = O(n^{-4/5} (\log n)^{3(2s-1)/5}),$$

where the multiplicative constant underlying $O(\cdot)$ depends on ρ, d, σ, and V. In the univariate case, we can deduce from [68, Theorem 2.1] that

$$R_F(\hat{f}_{n,V}^{1,1}, f^*) \leq C_\rho \Big(\frac{\sigma^2 V^{1/2}}{n}\Big)^{4/5} + C_\rho \frac{\sigma^2}{n} \log n,$$

where C_ρ is a positive constant depending on ρ. In other words,

$$R_F(\hat{f}^{1,1}_{n,V}, f^*) = O(n^{-4/5}),$$

where the multiplicative constant underlying $O(\cdot)$ depends on ρ, σ, and V. Thus, what Theorem 2 tells us is that for general d and s, $\hat{f}^{d,s}_{n,V}$ can achieve the same rate $n^{-4/5}$, although it slightly deteriorates by a logarithmic factor depending on s. This suggests that our lasso-based MARS fitting method can avoid the curse of dimensionality to some extent and be a useful estimation technique in higher dimensions.

The key step of our proof of Theorem 2 is to find an upper bound on the metric entropy of $\mathcal{D}_m$ (under the L^2 norm), which is defined as the collection of all functions of the form

$$(x_1, \dots, x_m) \mapsto \int (x_1 - t_1)_+ \cdots (x_m - t_m)_+ \, d\nu(\mathbf{t}),$$

where $m \in [d]$ and ν is a signed measure on $[0,1]^m$ with total variation $|\nu|([0,1]^m) \le 1$. The following theorem contains our result on the metric entropy of $\mathcal{D}_m$.

Theorem 3. *There exist positive constants C_m and ϵ_m depending on m such that*

$$\log N(\epsilon, \mathcal{D}_m, \|\cdot\|_2) \le C_m \epsilon^{-1/2} \Big(\log \frac{1}{\epsilon}\Big)^{3(2m-1)/4}$$

for every $0 < \epsilon < \epsilon_m$. The logarithmic factor can be omitted when $m = 1$.

Remark 1. *If the class $\mathcal{D}_m$ is altered by replacing $(x-t)_+$ with $\mathbf{1}(x \ge t)$ and restricting ν to probability measures, one obtains the collection of all functions of the form*

$$(x_1, \dots, x_m) \mapsto \int \mathbf{1}(x_1 \ge t_1) \cdots \mathbf{1}(x_m \ge t_m) \, d\nu(\mathbf{t}) = \nu\Big(\prod_{j=1}^m [0, x_j]\Big).$$

This class of functions is indeed the collection of all probability distributions on $[0,1]^m$, whose upper bounds on the metric entropy were derived in [13]. Thus, we are basically extending the argument in [13] from $\mathbf{1}(x \ge t)$ to $(x-t)_+$.

Theorem 3 is novel to the best of our knowledge even though we use standard tools and techniques for proving it. We first connect upper bounds of the metric entropy of $\mathcal{D}_m$ to lower bounds of the small ball probability of integrated Brownian sheet based on ideas from [13, Section 3] and [56, Section 3] and results from [94, Theorem 1.2] and [3, Theorem 5]. The small ball probability of integrated Brownian sheet here refers to the quantity

$$\mathbb{P}\Big(\sup_{\mathbf{t} \in [0,1]^m} |X_m(\mathbf{t})| \le \epsilon\Big),$$

where $\epsilon > 0$ and X_m is an m-dimensional integrated Brownian sheet. Required bounds on this small ball probability are then obtained using results from [44, Theorem 6] and [30, Theorem 1.2]. Specifically, we show that there exist positive constants c_m and ϵ_m depending on m such that

$$\log \mathbb{P}\Big(\sup_{\mathbf{t} \in [0,1]^m} |X_m(\mathbf{t})| \le \epsilon \Big) \ge -c_m \epsilon^{-2/3} \Big(\log \frac{1}{\epsilon} \Big)^{2m-1}$$

for every $0 < \epsilon < \epsilon_m$. This result, along with the connection between the metric entropy and the small ball probability, leads to Theorem 3, which is the main ingredient in our proof of Theorem 2. We refer the reader to Appendix C.8 for detailed explanation of the terminology and arguments.

Now, we turn to the result for the approximate version $\widetilde{f}^{d,s}_{n,V}$. The next theorem presents an upper bound on the risk of $\widetilde{f}^{d,s}_{n,V}$ under the same assumption as in Theorem 2. Recall that N_j are the pre-selected integers for the approximate method.

Theorem 4. *Suppose $f^* \in \mathcal{F}^{d,s}_{\infty-mars}$ and $V_{mars}(f^*) \le V$. Also, assume the lattice design* (3.19) *and sub-Gaussian errors* (3.20). *The estimator $\widetilde{f}^{d,s}_{n,V}$ then satisfies*

$$R_F(\widetilde{f}^{d,s}_{n,V}, f^*) \le \frac{8V^2}{N^2} + C_{\rho,d} \Big(\frac{\sigma^2 V^{1/2}}{n} \Big)^{4/5} \Big[\log \Big(2 + \frac{V n^{1/2}}{\sigma} \Big) \Big]^{3(2s-1)/5} + C_{\rho,d} \frac{\sigma^2}{n} (\log n)^2$$

for some positive constant $C_{\rho,d}$ depending on ρ and d, where $N = \min_j N_j$.

Theorem 4 shows that $\widetilde{f}^{d,s}_{n,V}$ has almost the same risk upper bound as $\hat{f}^{d,s}_{n,V}$. The only difference is the existence of the approximation error term $8V^2/N^2$, which converges to 0 as N goes to infinity. Hence, for sufficiently large N, $\widetilde{f}^{d,s}_{n,V}$ achieves the same rate as $\hat{f}^{d,s}_{n,V}$. Indeed, if we set each N_j to be of order at least $n^{2/5}$, then

$$R_F(\widetilde{f}^{d,s}_{n,V}, f^*) = O(n^{-4/5} (\log n)^{3(2s-1)/5}) \tag{3.23}$$

where the multiplicative constant underlying $O(\cdot)$ depends on ρ, d, σ, and V.

Random Design

Here, we assume that $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ are i.i.d. random vectors with density p_0 on $[0,1]^d$ that is bounded by some positive constant B, i.e.,

$$0 \le p_0(\mathbf{x}) \le B \quad \text{for all } \mathbf{x} \in [0,1]^d. \tag{3.24}$$

Also, we assume that $y_1, \dots, y_n$ are generated according to the regression model

$$y_i = f^*(\mathbf{x}^{(i)}) + \xi_i \tag{3.25}$$

where ξ_i are i.i.d. errors independent of $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ with mean zero and with finite $L^{5,1}$ norm; that is,

$$\|\xi_i\|_{5,1} := \int_0^\infty (\mathbb{P}(|\xi_i| > t))^{1/5} \, dt < \infty. \tag{3.26}$$

Note that condition (3.26) is stronger than the finite fifth-moment condition $\|\xi_i\|_5 < \infty$, but weaker than the finite $(5+\epsilon)^{\text{th}}$-moment condition $\|\xi_i\|_{5+\epsilon} < \infty$ for every $\epsilon > 0$ (see, e.g., [62, Chapter 1.4]). In this setting, we measure the accuracy of an estimator $\hat{f}_n$ of f^* by

$$\|\hat{f}_n - f^*\|_{p_0,2}^2 := \int \left(\hat{f}_n(\mathbf{x}) - f^*(\mathbf{x})\right)^2 p_0(\mathbf{x})\, d\mathbf{x}. \tag{3.27}$$

The next theorem presents the rate of convergence of $\hat{f}_{n,V}^{d,s}$ under the assumption $f^* \in \mathcal{F}_{\infty-\text{mars}}^{d,s}$ and $V_{\text{mars}}(f^*) \leq V$. Note that $\hat{f}_{n,V}^{d,s}$ still achieves the rate $n^{-4/5}$ as in the fixed lattice design setting, although the exponent of the logarithmic factor is slightly bigger when $s > 2$.

Theorem 5. *Suppose $f^* \in \mathcal{F}_{\infty-mars}^{d,s}$ and $V_{mars}(f^*) \leq V$. Also, assume that the density p_0 satisfies* (3.24) *and that the error terms ξ_i satisfy* (3.26)*. We then have*

$$\|\hat{f}_{n,V}^{d,s} - f^*\|_{p_0,2}^2 = O_p(n^{-4/5}(\log n)^{8(s-1)/5}). \tag{3.28}$$

As the metric entropy of $\mathcal{D}_m$ played a central role in our proof of Theorem 2, the bracketing entropy of $\mathcal{D}_m$ is the key ingredient of our proof of Theorem 5. The following theorem states an upper bound on the bracketing entropy of $\mathcal{D}_m$.

Theorem 6. *There exists a positive constant C_m depending on m such that*

$$\log N_{[\,]}(\epsilon, \mathcal{D}_m, \|\cdot\|_2) \leq C_m \left(\frac{4}{\epsilon}\right)^{1/2} \left|\log \frac{4}{\epsilon}\right|^{2(m-1)}$$

for every $\epsilon > 0$, where $N_{[\,]}(\epsilon, \mathcal{D}_m, \|\cdot\|_2)$ is the ϵ-bracketing number of $\mathcal{D}_m$ under the L^2 norm.

Remark 2. *Theorem 6 also provides an upper bound on the metric entropy of $\mathcal{D}_m$ (under the L^2 norm). Since*

$$N(\epsilon, \mathcal{D}_m, \|\cdot\|_2) \leq N_{[\,]}(2\epsilon, \mathcal{D}_m, \|\cdot\|_2),$$

we can derive from Theorem 6 that

$$\log N(\epsilon, \mathcal{D}_m, \|\cdot\|_2) \leq C_m \left(\frac{2}{\epsilon}\right)^{1/2} \left|\log \frac{2}{\epsilon}\right|^{2(m-1)}$$

for every $\epsilon > 0$. However, this upper bound is weaker than the one we achieved in Theorem 3. Although it has the same order for ϵ, the exponent of the logarithmic factor is bigger. We can obtain from this result an upper bound on the risk of $\hat{f}_{n,V}^{d,s}$ under the fixed lattice design setting, but it will lead to a bound looser than the one in Theorem 2.

We can prove a similar result as in Theorem 5 for the approximate version $\widetilde{f}_{n,V}^{d,s}$. As we state in the following theorem, $\widetilde{f}_{n,V}^{d,s}$ achieves the same rate of convergence as $\hat{f}_{n,V}^{d,s}$ if $N_1, \ldots, N_d$ are sufficiently large. Together with (3.23), this result suggests that the approximate method with appropriately chosen $N_1, \ldots, N_d$ can be as accurate as the original method.

Theorem 7. *Suppose* $f^* \in \mathcal{F}^{d,s}_{\infty-mars}$ *and* $V_{mars}(f^*) \leq V$*. Also, assume that the density* p_0 *satisfies* (3.24) *and that the error terms* ξ_i *satisfy* (3.26)*. Moreover, assume that* $N = \min_j N_j = \Omega(n^{4/15})$*, i.e., there exists a positive constant* $c_{B,d,V}$ *possibly depending on* $B, d,$ *and* V *such that*

$$N \geq c_{B,d,V} \cdot n^{4/15}.$$

Then, the estimator $\widetilde{f}^{d,s}_{n,V}$ *satisfies*

$$\|\widetilde{f}^{d,s}_{n,V} - f^*\|^2_{p_0,2} = O_p(n^{-4/5}(\log n)^{8(s-1)/5}).$$

Our next result shows that the logarithmic factor in (3.28) cannot be completely removed in the minimax sense. Specifically, we bound the minimax risk

$$\mathfrak{M}^{d,s}_{n,V} = \inf_{\hat{f}_n} \sup_{\substack{f^* \in \mathcal{F}^{d,s}_{\infty-\text{mars}} \\ V_{\text{mars}}(f^*) \leq V}} \mathbb{E}_{f^*}\|\hat{f}_n - f^*\|^2_{p_0,2},$$

where the expectation is taken over $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$ of (3.25) and $\inf_{\hat{f}_n}$ denotes the infimum over all estimators $\hat{f}_n$ of f^* based on $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$. Here, we further restrict that ξ_i in the model (3.25) are independent Gaussian errors with mean zero and variance σ^2 and that the density p_0 is bounded below by some positive constant b, i.e.,

$$p_0(\mathbf{x}) \geq b \quad \text{for all } \mathbf{x} \in [0,1]^d. \tag{3.29}$$

Our result shows that the supremum risk of every estimator indeed requires a logarithmic factor depending on s in addition to the $n^{-4/5}$ term. Note though that there is still a gap between the exponent $8(s-1)/5$ of $\log n$ in the rate of convergence of $\hat{f}^{d,s}_{n,V}$ and the exponent $4(s-1)/5$ of $\log n$ in the minimax lower bound.

Theorem 8. *Suppose that the density* p_0 *satisfies* (3.24) *and* (3.29) *and that* $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$*. Then, there exist positive constants* $C_{b,B,s}$ *depending on* $b, B,$ *and* s *and* $c_{B,s}$ *depending on* B *and* s *such that*

$$\mathfrak{M}^{d,s}_{n,V} \geq C_{b,B,s}\Big(\frac{\sigma^2 V^{1/2}}{n}\Big)^{4/5}\Big[\log\Big(\frac{Vn^{1/2}}{\sigma}\Big)\Big]^{4(s-1)/5}$$

provided $n \geq c_{B,s}(\sigma^2/V^2)$*.*

Our proof of Theorem 8 is based on Assouad's lemma with a finite set of functions in $\{f^* \in \mathcal{F}^{d,s}_{\infty-\text{mars}} : V_{\text{mars}}(f^*) \leq V\}$ that is constructed by an extension of the ideas in [13, Section 4]. Results similar to Theorem 8 can be proved under the fixed-design setting, but we do not go into detail in this thesis.

3.4 Numerical Experiments

In this section, we provide the results of some numerical experiments illustrating the performance of our estimators from either the original or the approximate method. The performance of our estimators is compared to the performance of the usual MARS estimator in our experiments. The results for simulated data are presented first, and those for real data follow next. Our methods are implemented in the R package regmdc, which is available at `https://github.com/DohyeongKi/regmdc`. Our R package regmdc employs the R package Rmosek (based on interior point convex optimization) to solve the finite-dimensional lasso problem (3.16). For the usual MARS estimator, we use the R package earth based on [53] and [51]. The code for all the results in this section is available at `https://github.com/DohyeongKi/mars-lasso-paper`.

Simulation Studies

[100] and [145] demonstrated that their locally adaptive regression splines and trend filtering excel at estimating functions with locally varying smoothness. Considering that our estimator is a multivariate generalization of the piecewise locally adaptive regression splines estimator and the first-order trend filtering estimator, it is natural to examine whether our methods also excel at adapting to the local variation of functions. To test the local adaptivity of our methods and compare it to the local adaptivity of the usual MARS method, we exploit in our simulation studies the following four functions (Function 1, Function 2, Function 3, and Function 4) whose smoothness varies significantly over the domain.

The definitions of the four functions (Function 1, Function 2, Function 3, and Function 4) are presented below. For each function, we consider the uniform design where $\mathbf{x}^{(i)}$ are uniformly distributed on $[0,1]^d$ and y_i are generated according to the model (3.25). For Function 1 and Function 3, we also consider the equally spaced lattice design where

$$\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} = \prod_{j=1}^{d} \Big\{ \frac{l_j}{n_j} : l_j \in \{0, \dots, n_j - 1\} \Big\}$$

for some integers $n_j \geq 2$. In all cases, independent Gaussian errors with standard deviation 1 are added to function evaluations. Simulations are performed at various sample sizes for each function as presented in Table 3.1.

- *Function 1 (L1).* The first function $f^* : [0,1]^2 \to \mathbb{R}$ is defined by

  $$f^*(x_1, x_2) = 10 \exp(-5 \cdot r(x_1, x_2)) \cdot \cos(10\pi \cdot r(x_1, x_2))$$

  for $x_1, x_2 \in [0,1]$, where $r(x_1, x_2) = ((x_1 - 0.3)^2 + (x_2 - 0.4)^2)^{1/2}$. This function represents a two-dimensional damped sinusoidal wave.

- *Function 2 (L2).* The second function $f^* : [0,1]^5 \to \mathbb{R}$ is defined as in Function 1 but additionally includes three dummy variables x_3, x_4, and x_5.

- *Function 3 (L3).* The third function $f^* : [0,1]^2 \to \mathbb{R}$ is defined by

$$f^*(x_1, x_2) = 5 \sin\Big(\frac{4}{(x_1^2 + x_2^2)^{1/2} + 0.001}\Big) + 7.5$$

  for $x_1, x_2 \in [0,1]$. This is a (scaled) two-dimensional version of the Doppler function used for simulation studies in [100, 145]. We add 0.001 to avoid division by zero.

- *Function 4 (L4).* The fourth function $f^* : [0,1]^5 \to \mathbb{R}$ is defined as in Function 3 but additionally includes three dummy variables x_3, x_4, and x_5.

We also use for comparison the following four Friedman's functions (see [53, Sections 4.3 and 4.4]), which have been frequently utilized to measure the performance of nonparametric regression methods (see, e.g., [104, 121, 122]). The four functions (Function 5, Function 6, Function 7, and Function 8) are defined as below, and for each function, we generate data $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$ according to the model (3.25) where $\mathbf{x}^{(i)}$ are uniformly distributed on $[0,1]^d$. For Function 5 and Function 6, we consider Gaussian errors with standard deviation 1 as in [53, Section 4.3]. For Function 7 and Function 8, we consider Gaussian errors with standard deviation 125 and 0.1, which yield a 3:1 signal to noise ratio as designed in [53, Section 4.4] (see also [104]). Also, for each function, we conduct experiments on three different sample sizes as in [53]: 50, 100, and 200 for Function 5 and Function 6; 100, 200, and 400 for Function 7 and Function 8.

- *Function 5 (F1).* The first function $f^* : [0,1]^{10} \to \mathbb{R}$ is defined by

$$f^*(x_1, \ldots, x_{10}) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$$

  for $x_1, \ldots, x_{10} \in [0,1]$.

- *Function 6 (F2).* The second function $f^* : [0,1]^5 \to \mathbb{R}$ is defined as in Function 5, but here the five variables $x_6, \ldots, x_{10}$ are removed.

- *Function 7 (F3).* The third function $f^* : [0,1]^4 \to \mathbb{R}$ is defined by

$$f^*(x_1, x_2, x_3, x_4) = \Big[(T_1(x_1))^2 + \big(T_2(x_2) \cdot x_3 - (T_2(x_2) \cdot T_4(x_4))^{-1}\big)^2\Big]^{1/2}$$

  for $x_1, x_2, x_3, x_4 \in [0,1]$, where T_1, T_2, and T_4 are the linear maps defined by $T_1(x_1) = 100x_1$, $T_2(x_2) = 2\pi(260x_2 + 20)$, and $T_4(x_4) = 10x_4 + 1$.

- *Function 8 (F4).* The fourth function $f^* : [0,1]^4 \to \mathbb{R}$ is defined by

$$f^*(x_1, x_2, x_3, x_4) = \arctan\Big(\frac{T_2(x_2) \cdot x_3 - (T_2(x_2) \cdot T_4(x_4))^{-1}}{T_1(x_1)}\Big)$$

  for $x_1, x_2, x_3, x_4 \in [0,1]$, where T_1, T_2, and T_4 are defined as in Function 7.

In all simulations, we compare the estimator $\hat{f}^{d,2}_{n,V}$ and its approximate version $\widetilde{f}^{d,2}_{n,V}$ to the usual MARS estimator whose order of interaction is restricted to 2. The approximate method is considered only when covariates are uniformly distributed. The positive integers N_j for the approximate method are set to 25 for each coordinate. Also, the tuning parameter V is selected by 10-fold cross-validation in all cases. For each function, we repeat the above data generation and estimator construction processes 25 times and average out computed losses. We use the squared empirical L^2 norm (3.21) for the fixed lattice design, and for the uniform design, we generate 1000 new samples and approximate the prediction error (3.27).

Table 3.1 presents the average loss of our estimators and the usual MARS estimator over 25 repetitions for each function. In Table 3.1, we first observe that our estimators outperform the usual MARS estimator in capturing the local variation of the first four functions, L1, L2, L3, and L4, under both the lattice and the uniform design. The only exceptions are when we estimate the function L2 with 100 samples and the function L4 with 100, 200, and 400 samples. Even for these functions, our methods start performing better when they are given more samples. It turns out that our methods usually benefit more from increases in sample sizes compared to the usual MARS method. By comparing L1 with L2 and L3 with L4, we can also see that our methods however suffer more from the addition of dummy variables.

Table 3.1 also shows that our methods estimate Friedman's functions F2, F3, and F4 much better, while the usual MARS method does better in the estimation of the function F1. However, the performance gap in estimating F1 narrows as the sample size increases, which reaffirms that increases in sample sizes are often more favorable for our methods than the usual MARS method. Also, comparing the results of F2 with those of F1, again we can see that our methods can benefit more from removing unnecessary covariates in advance. This hints that having appropriate variable selection as a pre-processing step can improve the performance of our methods. We think it can be a promising avenue for future work.

Moreover, we observe from Table 3.1 that the original and approximate method yield almost the same outputs in most cases. Only for the Doppler function L3, which has extremely wild fluctuation around the origin, the two methods show a notable difference. This tells us there may not be much degradation in practice from using the approximate method in place of the original one, and there even can be some gains, as in the case of L3.

From these simulation results, we can expect that although the usual MARS estimator might perform better on small sample sizes, our estimators usually outdo the usual MARS estimator when given enough samples. At what sample sizes such a transition occurs will definitely vary across functions, but in the above examples, they were reasonably small.

Real Datasets

Here, we use a few standard real datasets (Earnings, Airfoil Self-Noise, Abalone, Concrete, Ozone, Red Wine, and White Wine dataset) to compare our estimators with the usual MARS estimator. Brief descriptions of each dataset are presented in Appendix A.2.

For each dataset, we first linearly transform each covariate into $[0, 1]$. In general, there can be multiple options for linear transformations. If the domain of a covariate is known

Table 3.1: The average and standard error of squared empirical L^2 norms (3.21) (lattice design) or prediction errors (3.27) (uniform design) for each experimental setting. The number below the name of each function (if any) is the scale that should be multiplied to the corresponding averages and standard errors.

Function	Design	# of Data	Usual	Original	Approximate
L1	Lattice	256	3.66 (0.07)	**1.06 (0.07)**	–
		1024	3.34 (0.04)	**1.09 (0.05)**	–
	Uniform	100	3.30 (0.15)	3.21 (0.09)	**3.20 (0.08)**
		200	3.18 (0.09)	**3.00 (0.05)**	3.02 (0.05)
		400	2.90 (0.05)	**2.79 (0.05)**	2.80 (0.04)
		800	2.90 (0.03)	–	**2.62 (0.03)**
L2	Uniform	100	**3.29 (0.08)**	3.59 (0.07)	3.59 (0.07)
		200	3.15 (0.08)	**3.13 (0.05)**	**3.13 (0.05)**
		400	3.12 (0.05)	**3.09 (0.04)**	**3.09 (0.04)**
		800	2.92 (0.06)	–	**2.91 (0.05)**
L3	Lattice	256	3.76 (0.08)	**1.83 (0.07)**	–
		1024	3.43 (0.07)	**1.57 (0.05)**	–
	Uniform	100	5.37 (0.16)	5.20 (0.23)	**5.11 (0.21)**
		200	4.06 (0.15)	3.59 (0.08)	**3.43 (0.09)**
		400	3.54 (0.07)	3.06 (0.17)	**2.27 (0.08)**
		800	3.28 (0.07)	–	**1.51 (0.06)**
L4	Uniform	100	**6.25 (0.23)**	7.62 (0.14)	7.62 (0.14)
		200	**4.47 (0.17)**	6.03 (0.12)	5.94 (0.10)
		400	**3.63 (0.07)**	4.30 (0.08)	4.25 (0.09)
		800	3.28 (0.05)	–	**3.10 (0.06)**
F1	Uniform	50	**4.05 (0.37)**	8.60 (0.35)	8.89 (0.36)
		100	**1.03 (0.09)**	4.39 (0.36)	4.49 (0.38)
		200	**0.43 (0.02)**	0.87 (0.03)	0.87 (0.03)
F2	Uniform	50	3.25 (0.28)	**2.45 (0.31)**	2.48 (0.33)
		100	1.13 (0.10)	0.77 (0.12)	**0.72 (0.07)**
		200	0.41 (0.03)	0.37 (0.03)	**0.35 (0.02)**
F3 ($\cdot 10^3$)	Uniform	100	3.62 (0.37)	**2.10 (0.23)**	**2.10 (0.23)**
		200	2.16 (0.22)	**0.95 (0.07)**	**0.95 (0.07)**
		400	1.20 (0.15)	**0.55 (0.04)**	**0.55 (0.04)**
F4 ($\cdot 10^{-3}$)	Uniform	100	14.0 (1.04)	**11.5 (0.60)**	11.6 (0.63)
		200	9.12 (0.78)	**7.40 (0.34)**	7.43 (0.34)
		400	5.51 (0.35)	5.25 (0.24)	**5.23 (0.23)**

Table 3.2: The number of covariates, the number of data, and the type of our method used for each dataset. Original and Approx stand for the original and approximate method.

Dataset	Dimension	# of Data	Method
Earnings	2	25437	Original
Airfoil Self-Noise	5	1503	Original
Abalone	7	4177	Approx
Concrete	8	1030	Approx
Ozone	9	330	Approx
Red Wine	11	1599	Approx
White Wine	11	4898	Approx

as $[m, M]$, then it is natural to subtract m from it and then divide it by $M - m$. In case the two extreme values of the domain, m and M, are unknown, we can consider the same linear transformation after simply setting m as the minimum and M as the maximum among observed values. Here, we choose the latter option for all datasets.

We also split each dataset into a training set and a test set and use the mean squared error on the test set for comparison. We use 80% observations as training data and the remaining 20% observations as test data. Also, for every dataset, we focus on interactions between covariates of order up to 2. In other words, we use $\hat{f}^{d,2}_{n,V}$ or its approximate version $\widetilde{f}^{d,2}_{n,V}$ for estimating regression functions and compare it to the usual MARS estimator whose order of interaction is restricted to 2. For the Earnings and the Airfoil Self-Noise datasets, we employ the original method, and for the other datasets, we employ the approximate method, with setting each N_j to 25. Furthermore, the tuning parameter V is chosen by 10-fold cross-validation in all cases.

In Table 3.2, we present for each dataset the number of covariates, the number of data, and the type of our method we use (whether the original or the approximate method). Table 3.3 shows the average mean squared error of our estimator and the usual MARS estimator over 25 random training and test set splits for each dataset. In Table 3.3, we can see that our method outperforms the usual MARS method in almost all examples, while showing great improvement on the Airfoil Self-Noise, Concrete, Ozone, and White Wine dataset. The Red Wine dataset is the only exception for which the usual MARS method produces better or equally good fits. However, for its sibling (the White Wine dataset), which has about three times as many observations, our method is clearly a better option than the usual MARS method. This again proves that the usual MARS method may beat our method on small sample sizes, but our method reclaims the throne once enough data are provided.

We can conclude from all the results in this section that not only can our estimators be an alternative to the usual MARS estimator, but also they can supplement each other.

Table 3.3: The average and standard error of mean squared errors on test sets for each dataset. The number next to the name of each dataset is the scale that should be multiplied to the corresponding average and standard error.

Dataset	Usual Method	Our Method
Earnings ($\cdot 10^{-1}$)	2.68 (0.01)	**2.65 (0.01)**
Airfoil Self-Noise ($\cdot 10^{0}$)	9.16 (0.31)	**3.67 (0.13)**
Abalone ($\cdot 10^{0}$)	4.64 (0.08)	**4.59 (0.08)**
Concrete ($\cdot 10^{1}$)	4.01 (0.08)	**2.57 (0.17)**
Ozone ($\cdot 10^{1}$)	1.68 (0.08)	**1.45 (0.06)**
Red Wine ($\cdot 10^{-1}$)	**4.17 (0.08)**	**4.17 (0.07)**
White Wine ($\cdot 10^{-1}$)	5.19 (0.05)	**4.92 (0.07)**

Chapter 4

Totally Concave Regression

As explained in the Introduction, totally concave regression is a sign-constrained variant of the infinite-dimensional lasso framework (1.5) with the MARS atoms (3.2). The function class for totally concave regression is obtained by imposing sign constraints on the measures ν_S appearing in the representation (3.3). Specifically, it consists of all functions of the form

$$f_{\mathbf{c},\boldsymbol{\nu}}(x_1, \dots, x_d) := c_0 + \sum_{0<|S|\leq s} c_S \prod_{j\in S} x_j - \sum_{0<|S|\leq s} \int_{[0,1)^{|S|}\setminus\{\mathbf{0}\}} \prod_{j\in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S), \tag{4.1}$$

where c_0 and c_S are real numbers, and ν_S is a finite (Borel) measure on $[0,1)^{|S|} \setminus \{\mathbf{0}\}$ for each nonempty subset $S \subseteq [d]$ with $|S| \leq s$. As in (3.3), $\mathbf{0} := (0, \dots, 0)$, $(\cdot)_+ := \max(\cdot, 0)$, $|S|$ denotes the cardinality of S, and both summations are over all $S \subseteq [d] := \{1, \dots, d\}$ satisfying $0 < |S| \leq s$. Also, as in (3.3), the domain of these functions is taken to be $[0,1]^d$. In practice, this can be achieved by normalizing each covariate to lie in $[0,1]$ by subtracting its smallest possible value and then dividing by its range. Let $\mathcal{F}^{d,s}_{\mathrm{tc}}$ denote the class of all such functions. Clearly, $\mathcal{F}^{d,s}_{\mathrm{tc}}$ is a subclass of the MARS class $\mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ by construction.

The negative sign in front of the integral terms corresponds to the concave case. If this sign is replaced by a positive sign, one obtains the corresponding function class for totally convex regression. In this chapter, we focus on the concave case, although all our ideas apply to the convex setting as well. Unlike the MARS representation, where the multi-affine terms arise as the masses of the signed measures ν_S at $\mathbf{0}$, these terms are separated out from the measures and written explicitly in (4.1). Nevertheless, as in the MARS setting, the coefficients of these multi-affine terms remain unconstrained (recall (3.4)).

In this chapter, we study the function class $\mathcal{F}^{d,s}_{\mathrm{tc}}$ and use it for multivariate concave regression. We first explain how this function class can be naturally motivated for multivariate concave regression, independently of how we initially constructed it from the MARS framework. We begin by recalling univariate concave regression.

Consider a response variable y and a single covariate x, for which we observe data $(x^{(1)}, y_1), \dots, (x^{(n)}, y_n)$, where $x^{(i)} \in [0,1]$ and $y_i \in \mathbb{R}$. Univariate concave regression [73, 72] finds the least squares fit over all concave functions on $[0,1]$:

$$\hat{f}^1_{\text{concave}} \in \operatorname*{argmin}_{f \text{ is concave}} \sum_{i=1}^{n} \big(y_i - f(x^{(i)})\big)^2. \tag{4.2}$$

The estimator $\hat{f}^1_{\text{concave}}$ (where 1 denotes the covariate dimension) is well suited to settings with diminishing returns, in which the rate of change of y with respect to x decreases as x increases. This estimator is transparent (relying only on concavity), tuning-free, interpretable (yielding continuous piecewise affine fits), and enjoys strong theoretical guarantees under standard regression models when the true regression function is concave (see, e.g., [42, 64, 67]).

Our goal is to extend $\hat{f}^1_{\text{concave}}$ to multivariate regression settings where, instead of a single covariate x, one has d covariates $x_1, \dots, x_d$. We assume each x_j takes values in $[0,1]$ as before. The observations are $(\mathbf{x}^{(1)}, y_1), \dots, (\mathbf{x}^{(n)}, y_n)$, where $\mathbf{x}^{(i)} \in [0,1]^d$ and $y_i \in \mathbb{R}$. A natural approach to constructing multivariate extensions is to use the least squares estimator over a class of functions satisfying some multivariate notion of concavity. Three existing notions of multivariate concavity that are not the focus of this chapter are the following:

- **General concavity**: This is the usual multivariate notion of concavity corresponding to the condition

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \geq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) \quad \text{for all } \alpha \in [0,1] \text{ and } \mathbf{x}, \mathbf{y} \in [0,1]^d. \tag{4.3}$$

  Equivalently, f is generally concave if its restriction to every line segment $[\mathbf{x}, \mathbf{y}] := \{(1-\alpha)\mathbf{x} + \alpha\mathbf{y} : 0 \leq \alpha \leq 1\}$ is a univariate concave function. Nonparametric regression under general concavity has been studied in [87, 131, 98, 5]. In practice, there could be two issues with general concavity: (a) the assumption (4.3) may be difficult to justify in some applications, since it is required to hold for every $\mathbf{x}, \mathbf{y} \in [0,1]^d$; (b) the class of all generally concave functions on $[0,1]^d$ is large in the sense that its complexity grows exponentially with the dimension d. For example, the ϵ-metric entropy of uniformly bounded concave functions on $[0,1]^d$ grows as $\epsilon^{-d/2}$ (see, e.g., [65, 57]), which is exponential in d. As a result, the multivariate (generally) concave least squares estimator can suffer from overfitting due to the curse of dimensionality (in the sense of slow rates of convergence [70, 89]), limiting its use when d is moderate or large.

- **Axial concavity**: This refers to functions that are concave when restricted to any line parallel to one of the coordinate axes. Formally, we say that f is axially concave on $[0,1]^d$ if, for every $j \in [d]$ and every fixed $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d) \in [0,1]^{d-1}$, the function $g_j(t) := f(x_1, \dots, x_{j-1}, t, x_{j+1}, \dots, x_d)$ is concave in t over $[0,1]$. Axial concavity is also known as coordinate-wise concavity. Clearly, axial concavity is a weaker notion than general concavity, since it requires (4.3) only when $\mathbf{x}$ and $\mathbf{y}$ differ in at most one coordinate. Because of this weaker requirement, axial concavity may be easier to justify in practice than general concavity. On the other hand, the class of axially concave functions is very large, so overfitting can severely affect least squares regression procedures based solely on axial concavity. Perhaps, for this reason, axial concavity has received limited attention so far, with the exception of [75].

- **Additive concavity**: This refers to functions of the form

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{d} f_j(x_j) \tag{4.4}$$

where each f_j is a univariate concave function. Additive regression models with such constraints have been studied in [106, 124, 31, 105]. The resulting estimators are transparent and interpretable, require no tuning parameters, and enjoy strong theoretical guarantees. However, additive models are well known to be too restrictive for many applications, as they allow for no interaction effects.

Totally concave functions are obtained by augmenting additive concave functions with simple axially concave interaction terms. To formulate interaction terms for additive concave functions, we first recall the basic representation theorem for univariate concave functions (see, e.g., [110, Section 1.6]). It states that, for every concave function $g : [0, 1] \to \mathbb{R}$ satisfying the mild boundary regularity conditions,

$$g'(0+) := \lim_{t \to 0+} \frac{g(t) - g(0)}{t} < \infty \quad \text{and} \quad g'(1-) := \lim_{t \to 1-} \frac{g(1) - g(t)}{1 - t} > -\infty, \tag{4.5}$$

there exist unique $c_0, c_1 \in \mathbb{R}$ and a finite (Borel) measure ν on $(0, 1)$ such that

$$g(x) = c_0 + c_1 x - \int_{(0,1)} (x - t)_+ \, d\nu(t) \quad \text{for all } x \in [0, 1]. \tag{4.6}$$

Intuitively, (4.6) states that every concave function on $[0, 1]$ satisfying (4.5) can be written as an infinite linear combination of the atoms $x \mapsto -(x - t)_+$, $t \in (0, 1)$, together with affine terms. The regularity condition (4.5) is essential for the validity of (4.6) and is satisfied by all functions considered in this chapter.

Using the representation (4.6) for each component, we can write additive concave functions (4.4) as

$$f(x_1, \ldots, x_d) = c_0 + \sum_{j=1}^{d} c_j x_j - \sum_{j=1}^{d} \int_{(0,1)} (x_j - t_j)_+ \, d\nu_j(t_j) \tag{4.7}$$

where each ν_j is a finite measure on $(0, 1)$ and $c_0, \ldots, c_d$ are real coefficients. The right-hand side of (4.7) can be viewed as a linear model in the covariates $x_1, \ldots, x_d$ (with coefficients $c_1, \ldots, c_d$) and the transformed covariates $-(x_j - t_j)_+$ for $t_j \in (0, 1)$ (with coefficients represented by the measures ν_j).

We now augment (4.7) with axially concave interaction terms, leading to total concavity. Since (4.7) can be viewed as a linear model, it is natural to introduce interactions by taking products of the variables appearing in (4.7). When $d = 2$, there are four types of such product terms: x_1x_2, $(x_1 - t_1)_+x_2$ for $t_1 \in (0, 1)$, $x_1(x_2 - t_2)_+$ for $t_2 \in (0, 1)$, and $(x_1 - t_1)_+(x_2 - t_2)_+$ for $t_1, t_2 \in (0, 1)$. Appending these terms to the right-hand side of (4.7) yields the following modified model for $d = 2$:

$$\begin{aligned} f(x_1, x_2) = c_0 + c_1x_1 + c_2x_2 &- \int_{(0,1)} (x_1 - t_1)_+ \, d\nu_1(t_1) - \int_{(0,1)} (x_2 - t_2)_+ \, d\nu_2(t_2) \\ &+ c_{12}x_1x_2 - \int_{(0,1)} (x_1 - t_1)_+x_2 \, d\widetilde{\nu}_1(t_1) - \int_{(0,1)} x_1(x_2 - t_2)_+ \, d\widetilde{\nu}_2(t_2) \\ &- \int_{(0,1)^2} (x_1 - t_1)_+(x_2 - t_2)_+ \, d\nu_{12}(t_1, t_2), \end{aligned} \tag{4.8}$$

where $c_{12} \in \mathbb{R}$ and $\widetilde{\nu}_1$, $\widetilde{\nu}_2$, and ν_{12} are finite *signed* measures on $(0, 1)$, $(0, 1)$, and $(0, 1)^2$, respectively, representing the coefficients of the interaction terms. Total concavity arises by imposing *nonnegativity* on the signed measures $\widetilde{\nu}_1$, $\widetilde{\nu}_2$, ν_{12}, so that they simply become measures. This constraint ensures that each interaction term in (4.8) satisfies axial concavity, since each of the functions $(x_1, x_2) \mapsto -c(x_1 - t_1)_+x_2$, $(x_1, x_2) \mapsto -cx_1(x_2 - t_2)_+$, and $(x_1, x_2) \mapsto -c(x_1 - t_1)_+(x_2 - t_2)_+$ is axially concave if and only if $c \geq 0$.

The representation (4.8) can also be simplified further. The term $(x_1 - t_1)_+(x_2 - t_2)_+$ coincides with $x_1(x_2 - t_2)_+$ when $t_1 = 0$, and with $(x_1 - t_1)_+x_2$ when $t_2 = 0$. Thus, we can combine $\widetilde{\nu}_1$, $\widetilde{\nu}_2$, ν_{12} into a single measure on $[0, 1)^2 \setminus \{(0, 0)\}$, which leads to the following equivalent form:

$$\begin{aligned} f(x_1, x_2) = c_0 + c_1x_1 + c_2x_2 &- \int_{(0,1)} (x_1 - t_1)_+ \, d\nu_1(t_1) - \int_{(0,1)} (x_2 - t_2)_+ \, d\nu_2(t_2) \\ &+ c_{12}x_1x_2 - \int_{[0,1)^2 \setminus \{(0,0)\}} (x_1 - t_1)_+(x_2 - t_2)_+ \, d\nu_{12}(t_1, t_2), \end{aligned}$$

where $c_0, c_1, c_2, c_{12} \in \mathbb{R}$, and ν_1, ν_2, ν_{12} are finite measures on $(0, 1)$, $(0, 1)$, $[0, 1)^2 \setminus \{(0, 0)\}$.

For $d \geq 2$, totally concave functions are constructed in a similar fashion. By augmenting (4.7) with axially concave interaction terms as above, we obtain functions of the form:

$$f(x_1, \dots, x_d) = c_0 + \sum_{\emptyset \neq S \subseteq [d]} c_S \prod_{j \in S} x_j - \sum_{\emptyset \neq S \subseteq [d]} \int_{[0,1)^{|S|} \setminus \{\mathbf{0}\}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S), \tag{4.9}$$

where c_0 and c_S are real numbers, and ν_S is a finite (Borel) measure on $[0, 1)^{|S|} \setminus \{\mathbf{0}\}$ for each nonempty subset $S \subseteq [d]$. We denote by $\mathcal{F}^d_{\mathrm{tc}}$ the class of all functions of the form (4.9) and call it the class of totally concave functions on $[0, 1]^d$. Observe that $\mathcal{F}^d_{\mathrm{tc}}$ coincides with $\mathcal{F}^{d,s}_{\mathrm{tc}}$ in (4.1) with $s = d$; hence, $\mathcal{F}^d_{\mathrm{tc}} = \mathcal{F}^{d,d}_{\mathrm{tc}}$.

Functions of the form (4.9) are obtained by adding interaction terms of all orders to the additive model (4.7), together with sign constraints ensuring axial concavity of each interaction term. In practice, interactions of high order are rarely used, as they often provide limited additional flexibility while substantially increasing computational complexity. It is therefore natural to consider a subclass of $\mathcal{F}_{\text{tc}}^d$ in which only interactions of order at most s are included, for some pre-specified $s \in [d]$ (typically $s = 2$ or 3). After imposing this restriction, we arrive precisely at the class of functions in (4.1). Thus, $\mathcal{F}_{\text{tc}}^{d,s}$ is motivated not only as a sign-constrained variant of the MARS framework introduced earlier, but also independently as a structured extension of additive concave regression with controlled interaction effects.

In the classical references [120] and [54, Chapter 2], total concavity is defined in a different way, through the nonpositivity of certain divided differences. We recall this classical definition in Section 4.1 and prove that, under mild regularity conditions, it is equivalent to the measure-based definition introduced above. To the best of our knowledge, this equivalence has not been previously observed.

For sufficiently smooth functions, total concavity can also be characterized using derivatives, as is the case for other notions of multivariate concavity. Specifically, we prove that a smooth function f belongs to $\mathcal{F}_{\text{tc}}^d$ if and only if

$$f^{(p_1,\dots,p_d)} := \frac{\partial^{p_1+\cdots+p_d} f}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \le 0 \quad \text{for all } (p_1, \dots, p_d) \in \mathbb{Z}_{\ge 0}^d \text{ with } \max_j p_j = 2. \tag{4.10}$$

In words, total concavity corresponds to the nonpositivity of all mixed partial derivatives of maximum order two. By contrast, other notions of multivariate concavity are characterized by conditions on mixed partial derivatives with *total* order, $\sum_j p_j$, at most two (see Section 4.1). This highlights that total concavity differs fundamentally from other notions of multivariate concavity.

The goal of this chapter is to undertake a systematic study of the least squares estimator $\hat{f}_n^{d,s}$ over the class $\mathcal{F}_{\text{tc}}^{d,s}$ for each s, given by

$$\hat{f}_n^{d,s} \in \operatorname*{argmin}_{f \in \mathcal{F}_{\text{tc}}^{d,s}} \sum_{i=1}^n \left(y_i - f(\mathbf{x}^{(i)}) \right)^2, \tag{4.11}$$

and to explore its uses in practical regression problems. The estimators $\hat{f}_n^{d,s}$ with $2 \le s \le d$ are novel multivariate extensions of univariate concave regression. When $s = 1$, $\hat{f}_n^{d,s}$ reduces to the least squares estimator over the class of additive concave functions.

In Section 4.3, we study the rate of convergence of $\hat{f}_n^{d,s}$ for estimating the true regression function f^* under a standard fixed-design regression model. In particular, we show that, under the well-specified assumption $f^* \in \mathcal{F}_{\text{tc}}^{d,s}$, the rate of convergence is $n^{-4/5}(\log n)^{3(2s-1)/5}$ for $2 \le s \le d$ (see Theorem 9). This means that the usual curse of dimensionality is largely avoided by $\hat{f}_n^{d,s}$, even when s is taken as large as d. This relatively fast rate shows that total concavity is a promising shape constraint for multivariate regression. By contrast, rates of convergence for least squares estimators based on general or axial concavity are slower and more severely affected by the curse of dimensionality (see, e.g., [97, 6, 70, 89]).

Beyond achieving theoretical convergence rates that largely avoid the curse of dimensionality, we also provide evidence that totally concave regression is effective in practice.

In Section 4.4, we apply it to real-world regression problems and demonstrate competitive performance relative to other standard methods.

To improve practical prediction performance, we found it helpful to further regularize $\mathcal{F}_{\mathrm{tc}}^{d,s}$. Section 4.2 introduces a regularized variant that imposes an upper bound on the total masses of the measures defining totally concave functions. This approach, which combines total concavity with the smoothness constraint in Chapter 3, helps mitigate overfitting near the boundary of the covariate domain, a common issue with shape-constrained estimators. It also enables theoretical analysis of accuracy under random designs (see Section 4.3). For better performance when d is large, Section 4.2 also discusses hybrid approaches that impose total concavity on a subset of covariates while assuming linearity on the others. Their utility in data analysis is explored in Section 4.4.

The optimization problem (4.11) is infinite-dimensional, in the sense that it involves infinitely many parameters. However, as in the case of MARS, it can be reduced to a finite-dimensional convex optimization problem. Indeed, the same reduction idea as in Section 3.2 applies here. The resulting problem has the same form as (3.16), except that the L^1-norm constraint on the coefficients is replaced by sign constraints. Hence, it becomes a nonnegative least squares problem [136, 103] rather than a lasso problem. Moreover, as in the MARS setting, an approximate version of $\hat{f}_n^{d,s}$ can be obtained by using the proxy lattices (3.18) in place of the original lattices associated with the design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. Since the arguments from Section 3.2 extend to totally concave regression with only minor modifications, we omit a separate section on computation in this chapter. We refer the reader to [80, Section 4] for further details.

The results of this chapter are based on those in [80]. We refer the reader to Appendix D for the proofs of all results presented in this chapter. We only note here that, although many of the results in this chapter closely resemble those in Chapter 3, the proof techniques differ substantially because of a key distinction: in totally concave regression, the total masses of the measures are unconstrained, whereas in the MARS setting they are constrained. This parallels the relationship between lasso and nonnegative least squares in finite-dimensional regression, where similar results often arise from different assumptions and proofs.

4.1 Characterizations

In this section, we introduce the classical definition of total concavity, originally due to [120] (see also [54, Chapter 2]), which is formulated through the nonpositivity of certain divided differences. We refer to this classical notion as total concavity in the sense of Popoviciu in this chapter. We show that, under mild regularity conditions, this definition is equivalent to the measure-based definition of total concavity introduced above. We also establish the alternative characterization (4.10) for smooth functions, which expresses total concavity in terms of mixed partial derivatives of maximum order two. This characterization will allow us to clarify how total concavity differs from other notions of multivariate concavity.

We begin by recalling the divided differences needed for the definition of total concavity in the sense of Popoviciu.

Divided Differences

Divided differences appear frequently in numerical analysis (see, e.g., [74]). For multivariate functions $f : [0,1]^d \to \mathbb{R}$, there are several notions of divided differences. In this thesis, we work with the following "tensor-product" form of divided differences; further background can be found in the standard references [54, Chapter 2], [74, Section 6.6], [48, Section 3].

Definition 7 (Divided differences)**.** *Let f be a real-valued function on $[0,1]^d$, and fix non-negative integers $p_1, \dots, p_d$. For each $j \in [d]$, let $0 \le v_1^{(j)} < \cdots < v_{p_j+1}^{(j)} \le 1$ be $p_j + 1$ distinct points. The divided difference of f of order $(p_1, \dots, p_d)$ with respect to the points $(v_{i_j}^{(j)}, 1 \le i_j \le p_j + 1, 1 \le j \le d)$ is defined by*

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & ;\ f \\ v_1^{(d)}, \dots, v_{p_d+1}^{(d)} & \end{bmatrix} := \sum_{i_1=1}^{p_1+1} \cdots \sum_{i_d=1}^{p_d+1} \frac{f(v_{i_1}^{(1)}, \dots, v_{i_d}^{(d)})}{\prod_{l_1 \neq i_1}(v_{i_1}^{(1)} - v_{l_1}^{(1)}) \times \cdots \times \prod_{l_d \neq i_d}(v_{i_d}^{(d)} - v_{l_d}^{(d)})}. \tag{4.12}$$

Divided differences of the form (4.12) can be viewed as discrete analogues of mixed partial derivatives of f. Indeed, if f is sufficiently smooth and $(v^{(1)}, \dots, v^{(d)}) \in (0,1)^d$, then

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & ;\ f \\ v_1^{(d)}, \dots, v_{p_d+1}^{(d)} & \end{bmatrix} \to \frac{f^{(p_1,\dots,p_d)}(v^{(1)}, \dots, v^{(d)})}{p_1! \cdots p_d!}$$

provided that $v_{i_j}^{(j)} \to v^{(j)}$ for each $i_j = 1, \dots, p_j + 1$ and $j \in [d]$ (see, e.g., [48, Section 3]).

Total Concavity in the sense of Popoviciu

Total concavity in the sense of Popoviciu is defined as follows.

Definition 8 (Total concavity in the sense of Popoviciu)**.** *We say $f : [0,1]^d \to \mathbb{R}$ is totally concave in the sense of Popoviciu if, for every $(p_1, \dots, p_d) \in \mathbb{Z}_{\ge 0}^d$ with $\max_j p_j = 2$, we have*

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & ;\ f \\ v_1^{(d)}, \dots, v_{p_d+1}^{(d)} & \end{bmatrix} = \sum_{i_1=1}^{p_1+1} \cdots \sum_{i_d=1}^{p_d+1} \frac{f(v_{i_1}^{(1)}, \dots, v_{i_d}^{(d)})}{\prod_{l_1 \neq i_1}(v_{i_1}^{(1)} - v_{l_1}^{(1)}) \times \cdots \times \prod_{l_d \neq i_d}(v_{i_d}^{(d)} - v_{l_d}^{(d)})} \le 0 \tag{4.13}$$

for every $0 \le v_1^{(j)} < \cdots < v_{p_j+1}^{(j)} \le 1, j \in [d]$.

In other words, f is totally concave in the sense of Popoviciu if all divided differences of f of order $(p_1, \dots, p_d)$ are nonpositive whenever $\max_j p_j = 2$. We denote by $\mathcal{F}^d_{\mathrm{tcp}}$ the class of all functions on $[0,1]^d$ that are totally concave in the sense of Popoviciu.

Example 1 $(d = 2)$**.** *A function $f : [0,1]^2 \to \mathbb{R}$ is totally concave in the sense of Popoviciu (that is, $f \in \mathcal{F}^2_{tcp}$) if the divided differences of f of orders*

$$(p_1, p_2) = (2,0), (0,2), (2,1), (1,2), (2,2)$$

are all nonpositive.

The nonpositivity of the divided difference of f of order $(2,0)$ means that, for every $0 \le v_1 < v_2 < v_3 \le 1$ and $0 \le w \le 1$,

$$\begin{bmatrix} v_1, v_2, v_3 & ; & f \\ w & & \end{bmatrix} = \frac{f(v_1, w)}{(v_1 - v_2)(v_1 - v_3)} + \frac{f(v_2, w)}{(v_2 - v_3)(v_2 - v_1)} + \frac{f(v_3, w)}{(v_3 - v_1)(v_3 - v_2)} \le 0.$$

This can be rewritten as

$$\frac{f(v_3, w) - f(v_2, w)}{v_3 - v_2} \le \frac{f(v_2, w) - f(v_1, w)}{v_2 - v_1}.$$

Hence, the nonpositivity of the divided difference of f of order $(2,0)$ is equivalent to concavity of $f(\cdot, x_2)$ for each fixed x_2. Similarly, the nonpositivity of the divided difference of f of order $(0,2)$ is equivalent to concavity of $f(x_1, \cdot)$ for each fixed x_1. Therefore, f is axially concave if and only if the divided differences of f of orders $(2,0)$ and $(0,2)$ are nonpositive.

Next, the divided difference of f of order $(2,1)$ is nonpositive if and only if, for every $0 \le v_1 < v_2 < v_3 \le 1$ and $0 \le w_1 < w_2 \le 1$,

$$\frac{f(v_3, w_2) - f(v_2, w_2) - f(v_3, w_1) + f(v_2, w_1)}{(v_3 - v_2)(w_2 - w_1)} \le \frac{f(v_2, w_2) - f(v_1, w_2) - f(v_2, w_1) + f(v_1, w_1)}{(v_2 - v_1)(w_2 - w_1)}.$$

The corresponding condition for order $(1,2)$ is obtained by interchanging the roles of the two coordinates.

Lastly, the divided difference of f of order $(2,2)$ is nonpositive if and only if

$$\frac{f(v_3, w_3) - f(v_2, w_3) - f(v_3, w_2) + f(v_2, w_2)}{(v_3 - v_2)(w_3 - w_2)} - \frac{f(v_2, w_3) - f(v_1, w_3) - f(v_2, w_2) + f(v_1, w_2)}{(v_2 - v_1)(w_3 - w_2)}$$
$$- \frac{f(v_3, w_2) - f(v_2, w_2) - f(v_3, w_1) + f(v_2, w_1)}{(v_3 - v_2)(w_2 - w_1)} + \frac{f(v_2, w_2) - f(v_1, w_2) - f(v_2, w_1) + f(v_1, w_1)}{(v_2 - v_1)(w_2 - w_1)}$$

is nonpositive for every $0 \le v_1 < v_2 < v_3 \le 1$ and $0 \le w_1 < w_2 < w_3 \le 1$.

Remark 3 (Relation to axial concavity)**.** *Suppose f is a real-valued function on $[0,1]^d$. We saw in the preceding example that, when $d = 2$, f is axially concave if and only if the divided* *differences of f of orders $(2,0)$ and $(0,2)$ are nonpositive. More generally, for arbitrary d, f is axially concave if and only if the divided differences of f of orders*

$$(2,0,\dots,0),(0,2,0,\dots,0),\dots,(0,\dots,0,2)$$

are nonpositive. Compared with the condition $f \in \mathcal{F}^d_{tcp}$, this requires a strictly smaller collection of inequalities. Hence, axial concavity is a weaker notion of concavity than total concavity in the sense of Popoviciu.

Equivalence between the Two Definitions of Total Concavity

The two definitions of total concavity are equivalent under minor regularity conditions:

Proposition 4. *Every function $f \in \mathcal{F}^d_{tc}$ is totally concave in the sense of Popoviciu, that is, $f \in \mathcal{F}^d_{tcp}$. Conversely, every function $f \in \mathcal{F}^d_{tcp}$ satisfying the following additional regularity conditions belongs to $\mathcal{F}^d_{tc}$: for every nonempty $S \subseteq [d]$,*

$$\lim_{t \to 0+} \begin{bmatrix} 0, t, & j \in S & \\ 0, & j \notin S & ; \; f \end{bmatrix} < \infty \tag{4.14}$$

and

$$\lim_{t \to 1-} \begin{bmatrix} t, 1, & j \in S & \\ 0, & j \notin S & ; \; f \end{bmatrix} > -\infty. \tag{4.15}$$

Conditions (4.14) and (4.15) reduce to (4.5) when $d = 1$, so that Proposition 4 reduces to the univariate representation theorem (4.6).

Totally concave functions with restricted interactions, namely functions in $\mathcal{F}^{d,s}_{\mathrm{tc}}$ (see (4.1)), can also be described as totally concave functions in the sense of Popoviciu satisfying certain conditions. In addition to the regularity conditions above, one only needs a condition enforcing the interaction order restriction, which was previously achieved by simply dropping the measures corresponding to higher-order interactions in the representation (4.9). Here, this restriction can instead be imposed through the condition that

$$\begin{bmatrix} v_j, w_j, & j \in S & \\ 0, & j \notin S & ; \; f \end{bmatrix} = 0 \quad \text{for every } 0 \le v_j < w_j \le 1, j \in S \tag{4.16}$$

for every $S \subseteq [d]$ with $|S| > s$. We refer to (4.16) as the interaction restriction condition. It can also be verified by induction that this condition is equivalent to requiring that, for every subset $S \subseteq [d]$ with $|S| > s$,

$$\begin{bmatrix} v_j, w_j, & j \in S & \\ v_j, & j \notin S & ; \; f \end{bmatrix} = 0 \tag{4.17}$$

for every $0 \le v_j < w_j \le 1, j \in S$, and every $v_j \in [0,1], j \notin S$.

Proposition 5. *The function class $\mathcal{F}_{tc}^{d,s}$ consists precisely of all $f \in \mathcal{F}_{tcp}^{d}$ satisfying* (4.14) *and* (4.15) *for every nonempty $S \subseteq [d]$, and satisfying* (4.16) *for every $S \subseteq [d]$ with $|S| > s$.*

It can also be shown that the mild regularity conditions (4.14) and (4.15) do not fundamentally change the least squares estimation procedure. Specifically, the following result shows that $\hat{f}_n^{d,s}$ also minimizes least squares over all functions $f \in \mathcal{F}_{\text{tcp}}^{d}$ satisfying the interaction restriction condition (4.16) for every $S \subseteq [d]$ with $|S| > s$.

Proposition 6. *Every minimizer $\hat{f}_n^{d,s}$ of* (4.11) *satisfies*

$$\hat{f}_n^{d,s} \in \operatorname*{argmin}_{f} \left\{ \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)}) \right)^2 : f \in \mathcal{F}_{tcp}^{d} \text{ and satisfies } (4.16) \right\}.$$

In particular, $\hat{f}_n^{d,s}$ also minimizes least squares over the entire class $\mathcal{F}_{tcp}^{d}$.

Note that the least squares criterion depends only on the values of functions at the design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. Proposition 6 is therefore a direct consequence of the following lemma, which states that, for every function $f \in \mathcal{F}_{\text{tcp}}^{d}$ satisfying (4.16) for every $S \subseteq [d]$ with $|S| > s$, there always exists an element of $\mathcal{F}_{\text{tc}}^{d,s}$ that is indistinguishable from f on $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$.

Lemma 3. *Suppose $f \in \mathcal{F}_{tcp}^{d}$ satisfies* (4.16) *for every $S \subseteq [d]$ with $|S| > s$. Then, there exists $g \in \mathcal{F}_{tc}^{d,s}$ such that $g(\mathbf{x}^{(i)}) = f(\mathbf{x}^{(i)})$ for all $i = 1, \dots, n$.*

Smoothness Characterization

The next result shows that, for smooth functions, total concavity is characterized by the nonpositivity of all mixed partial derivatives of maximum order two. This result can be viewed as a continuous analogue of the divided difference constraint in Definition 8. As in the smoothness characterization for MARS (Lemma 1), we impose assumptions stronger than necessary, since the precise minimal smoothness requirements are not of interest here.

Proposition 7. *Let $f : [0,1]^d \to \mathbb{R}$ be such that $f^{(\mathbf{p})}$ exists and is continuous for every $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1,2\}^d$. Then, $f \in \mathcal{F}_{tc}^{d}$ if $f^{(\mathbf{p})} \le 0$ for all $\mathbf{p} \in \{0,1,2\}^d$ with $\max_j p_j = 2$. If, in addition, $f^{(\mathbf{p})} = 0$ for all $\mathbf{p} \in \{0,1\}^d$ with $\sum_j p_j > s$, then $f \in \mathcal{F}_{tc}^{d,s}$.*

Comparison with Other Notions of Multivariate Concavity

Every totally concave function is axially concave by construction. It is also clear that every additive concave function is totally concave by the definition of total concavity. There is, however, no inclusion relationship between total concavity and general concavity. Indeed, there exist totally concave functions, such as $(x_1, x_2) \mapsto x_1 x_2$, that are not generally concave. Conversely, there also exist generally concave functions, such as $(x_1, x_2) \mapsto (x_1 + x_2 - 0.5)_+$, that are not totally concave. These relationships are summarized in Figure 4.1.

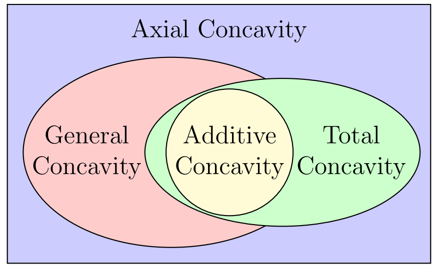

Figure 4.1: Four types of multivariate concavity.

For smooth functions f, these notions of concavity can be characterized as follows. Below, Hf denotes the Hessian of f.

- **General Concavity**: $-Hf \succcurlyeq 0$ (where $\succcurlyeq 0$ denotes positive semi-definiteness).
- **Axial Concavity**: $-\mathrm{diag}(Hf) \geq 0$ (where $\mathrm{diag}(Hf)$ denotes the diagonal of Hf, and ≥ 0 denotes element-wise nonnegativity).
- **Additive Concavity**: $-Hf = -\mathrm{diag}(Hf) \geq 0$.
- **Total Concavity**: $-f^{(p_1,\dots,p_d)} \geq 0$ for every $(p_1, \dots, p_d)$ with $\max_j p_j = 2$.

General, axial, and additive concavity are characterized via the Hessian matrix. Total concavity, by contrast, is fundamentally different: it involves higher-order derivatives and cannot be characterized via the Hessian.

4.2 Variants

Variants for Overfitting Reduction

Shape-constrained regression estimators tend to overfit near the boundary of the covariate domain, even in the univariate case (see, e.g., [153, 138, 86, 98, 102, 10, 95]). Totally concave regression exhibits a similar tendency. Indeed, the only restriction on the measures ν_S is nonnegativity, so they may place arbitrarily large mass near the boundary in order to achieve nearly perfect fits in that region. This issue can be mitigated by constraining the sum of the total masses $\nu_S([0,1)^{|S|} \setminus \{\mathbf{0}\})$:

$$\hat{f}^{d,s}_{n,V} \in \operatorname*{argmin}_f \left\{ \sum_{i=1}^n \left(y_i - f(\mathbf{x}^{(i)})\right)^2 : f \in \mathcal{F}^{d,s}_{\mathrm{tc}} \text{ and } V_{\mathrm{mars}}(f) \leq V \right\}, \tag{4.18}$$

where, for $f \in \mathcal{F}_{\text{tc}}^{d,s}$ of the form (4.1),

$$V_{\text{mars}}(f) := \sum_{0<|S|\leq s} \nu_S\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big), \tag{4.19}$$

and V is a tuning parameter, which can be chosen in practice by cross-validation.

We use the same notation $V_{\text{mars}}(\cdot)$ as in Chapter 3, since this is precisely the restriction of the MARS complexity measure to $\mathcal{F}_{\text{tc}}^{d,s}$. Because each ν_S is nonnegative, its variation $|\nu_S|$ coincides with ν_S. Thus, $\hat{f}_{n,V}^{d,s}$ can also be viewed as a constrained variant of the totally concave regression estimator $\hat{f}_n^{d,s}$, obtained by imposing an additional complexity constraint inherited from MARS. Estimators with analogous additional constraints, similar to $\hat{f}_{n,V}^{d,s}$, have been used to reduce boundary overfitting in the shape-constrained regression literature (see, e.g., [153, 138, 95]).

Variants for Handling Large d

The computational complexity of $\hat{f}_n^{d,s}$ increases with d. When d is large, it thus makes sense to assume total concavity only on a subset of the covariates, while imposing the stronger (and simpler) assumption of linearity on the remaining covariates. This leads to the model:

$$f^*(x_1, \dots, x_d) = f_1^*(x_1, \dots, x_p) + c_{p+1}x_{p+1} + \cdots + c_d x_d \tag{4.20}$$

where $f_1^* \in \mathcal{F}_{\text{tc}}^{p,s}$ and $c_{p+1}, \dots, c_d \in \mathbb{R}$. We denote this class by $\mathcal{F}_{\text{tc}}^{p,s}(x_1, \dots, x_p) \oplus \mathcal{L}(x_{p+1}, \dots, x_d)$ and the least squares estimator by $\hat{f}_n^{d,p,s}$. Using (4.1), we rewrite (4.20) as

$$c_0 + \sum_{\substack{S \subseteq [p] \\ 0<|S|\leq s}} c_S \prod_{j \in S} x_j - \sum_{\substack{S \subseteq [p] \\ 0<|S|\leq s}} \int_{[0,1)^{|S|}\setminus\{\mathbf{0}\}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S) + \sum_{j=p+1}^{d} c_j x_j.$$

A concrete example is given below.

Example 2 ($d = 4$, $p = 2$, and $s = 2$)**.** $\mathcal{F}_{tc}^{2,2}(x_1, x_2) \oplus \mathcal{L}(x_3, x_4)$ *consists of functions*

$$\begin{aligned} f^*(x_1, \dots, x_4) = c_0 + c_1 x_1 + c_2 x_2 - \int_{(0,1)} (x_1 - t_1)_+ \, d\nu_1(t_1) - \int_{(0,1)} (x_2 - t_2)_+ \, d\nu_2(t_2) \\ + c_{12} x_1 x_2 - \int_{[0,1)^2 \setminus \{(0,0)\}} (x_1 - t_1)_+ (x_2 - t_2)_+ \, d\nu_{12}(t_1, t_2) + c_3 x_3 + c_4 x_4, \end{aligned}$$

where $c_0, c_1, c_2, c_{12}, c_3, c_4 \in \mathbb{R}$*, and* ν_1, ν_2*, and* ν_{12} *are finite measures on* $(0,1)$*,* $(0,1)$*, and* $[0,1)^2 \setminus \{(0,0)\}$*, respectively.*

We further extend this model using interactions between covariates $x_1, \dots, x_p$ and a subset of the remaining covariates, $x_{p+1}, \dots, x_q$ (excluding $x_{q+1}, \dots, x_d$). These interactions take the form $\prod_{j\in S} x_j \cdot \prod_{k\in T} x_k$ and $\prod_{j\in S}(x_j - t_j)_+ \cdot \prod_{k\in T} x_k$, where $S \subseteq [p]$ and $T \subseteq [q]\setminus[p]$, with $|S \cup T| \leq s$ to maintain the interaction order limit. This results in functions

$$c_0 + \sum_{\substack{S\subseteq[p],T\subseteq[q]\setminus[p]\\0<|S\cup T|\leq s}} c_{S,T} \prod_{j\in S} x_j \cdot \prod_{k\in T} x_k$$
$$- \sum_{\substack{S\subseteq[p],T\subseteq[q]\setminus[p]\\|S|>0,|S\cup T|\leq s}} \int_{[0,1)^{|S|}\setminus\{\mathbf{0}\}} \prod_{j\in S}(x_j - t_j)_+ \cdot \prod_{k\in T} x_k \, d\nu_{S,T}(t_j, j\in S) + \sum_{j=q+1}^{d} c_j x_j.$$

We denote the collection of these functions by $\mathcal{F}_{\text{tc}}^{q,s}(x_1,\ldots,x_p; x_{p+1},\ldots,x_q) \oplus \mathcal{L}(x_{q+1},\ldots,x_d)$ and the corresponding least squares estimator by $\hat{f}_n^{d,q,p,s}$.

Example 3 ($d = 4$, $q = 3$, $p = 2$, and $s = 2$)**.** *$\mathcal{F}_{tc}^{3,2}(x_1, x_2; x_3) \oplus \mathcal{L}(x_4)$ is the collection of all functions of the form*

$$f^*(x_1,\ldots,x_4) = c_0 + c_1x_1 + c_2x_2 + c_3x_3 - \int_{(0,1)} (x_1 - t_1)_+ \, d\nu_1(t_1) - \int_{(0,1)} (x_2 - t_2)_+ \, d\nu_2(t_2)$$
$$+ c_{12}x_1x_2 - \int_{[0,1)^2\setminus\{(0,0)\}} (x_1 - t_1)_+(x_2 - t_2)_+ \, d\nu_{12}(t_1, t_2)$$
$$+ c_{13}x_1x_3 - \int_{(0,1)} (x_1 - t_1)_+ x_3 \, d\nu_{13}(t_1) + c_{23}x_2x_3 - \int_{(0,1)} (x_2 - t_2)_+ x_3 \, d\nu_{23}(t_2) + c_4x_4.$$

Similarly to (4.19), these functions can also be regularized by constraining the total masses of measures:

$$V_{\text{mars}}(f) := \sum_{\substack{S\subseteq[p],T\subseteq[q]\setminus[p]\\|S|>0,|S\cup T|\leq s}} \nu_{S,T}\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big).$$

The regularized variant $\hat{f}_{n,V}^{d,q,p,s}$ of $\hat{f}_n^{d,q,p,s}$ is then defined as the least squares estimator over the class of functions $f \in \mathcal{F}_{\text{tc}}^{q,s}(x_1,\ldots,x_p; x_{p+1},\ldots,x_q) \oplus \mathcal{L}(x_{q+1},\ldots,x_d)$ that satisfy the additional constraint $V_{\text{mars}}(f) \leq V$.

These function classes are substantially smaller than $\mathcal{F}_{\text{tc}}^{d,s}$ and thus provide stronger regularization, making them better suited for higher-dimensional regression problems.

4.3 Theoretical Accuracy

We study the theoretical accuracy of our estimators under the standard regression model:

$$y_i = f^*(\mathbf{x}^{(i)}) + \xi_i, \tag{4.21}$$

where f^* is the unknown regression function and ξ_i are mean-zero error terms. We consider both fixed-design and random-design settings.

Fixed Design

Here, we assume that $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ form the following equally-spaced lattice:

$$\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\} = \prod_{j=1}^{d} \Big\{0, \frac{1}{n_j}, \ldots, \frac{n_j - 1}{n_j}\Big\} \tag{4.22}$$

for some integers $n_j \geq 2$ for $j \in [d]$. We also assume that $y_1, \ldots, y_n$ are generated according to the model (4.21) where $\xi_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. For an estimator $\hat{f}_n$ of f^* based on the data $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$, we define the loss and risk of $\hat{f}_n$ by

$$\|\hat{f}_n - f^*\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \big(\hat{f}_n(\mathbf{x}^{(i)}) - f^*(\mathbf{x}^{(i)})\big)^2 \quad \text{and} \quad R_F(\hat{f}_n, f^*) := \mathbb{E}\|\hat{f}_n - f^*\|_n^2,$$

where the expectation is taken over $y_1, \ldots, y_n$.

The following theorem shows that the convergence rate of $\hat{f}_n^{d,s}$ is of order $n^{-4/5}$, up to logarithmic factors, when the true regression function f^* belongs to $\mathcal{F}_{\text{tc}}^{d,s}$. This result involves the quantity

$$V_{\text{design}}(f) := \inf \big\{V_{\text{mars}}(g) : g \in \mathcal{F}_{\text{tc}}^{d} \text{ such that } g(\mathbf{x}^{(i)}) = f(\mathbf{x}^{(i)}) \text{ for } i = 1, \ldots, n\big\},$$

where $V_{\text{mars}}(g)$ is as defined in (4.19) for $s = d$. By Lemma 3, for every $f \in \mathcal{F}_{\text{tcp}}^{d}$, there exists $g \in \mathcal{F}_{\text{tc}}^{d}$ that agrees with f at all design points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$. This ensures that $V_{\text{design}}(f)$ is finite for every $f \in \mathcal{F}_{\text{tcp}}^{d}$. The next lemma gives a simple bound on $V_{\text{design}}(f)$ for $f \in \mathcal{F}_{\text{tcp}}^{d}$.

Lemma 4. *If $f \in \mathcal{F}_{tcp}^{d}$, then*

$$V_{design}(f) \leq \sum_{\emptyset \neq S \subseteq [d]} \left(\begin{bmatrix} 0, 1/n_j, & j \in S & ; & f \\ 0, & j \notin S & & \end{bmatrix} - \begin{bmatrix} (n_j - 1)/n_j, 1, & j \in S & ; & f \\ 0, & j \notin S & & \end{bmatrix} \right).$$

We are now ready to state the main result of this section. Note that the risk bounds below depend on $V_{\text{design}}(f^*)$.

Theorem 9. *Suppose $f^* \in \mathcal{F}_{tcp}^{d}$ and satisfies (4.16) for every $S \subseteq [d]$ with $|S| > s$. Assume the equally-spaced lattice design (4.22) and $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. With $V := V_{design}(f^*)$, we have*

$$\begin{aligned} R_F(\hat{f}_n^{d,s}, f^*) \leq{}& C_d \Big(\frac{\sigma^2 V^{1/2}}{n}\Big)^{4/5} \Big[\log\Big(\Big(1 + \frac{V}{\sigma}\Big)n\Big)\Big]^{3(2s-1)/5} \\ &+ C_d \frac{\sigma^2}{n} (\log n)^{5d/4} \Big[\log\Big(\Big(1 + \frac{V}{\sigma}\Big)n\Big)\Big]^{3(2s-1)/4} \end{aligned}$$

if $s \geq 2$, and

$$R_F(\hat{f}_n^{d,1}, f^*) \leq C_d \Big(\frac{\sigma^2 V^{1/2}}{n}\Big)^{4/5} + C_d \frac{\sigma^2}{n} (\log n)^{d} \Big[\log\Big(\Big(1 + \frac{V}{\sigma}\Big)n\Big)\Big]^{2} + C_d \frac{\sigma^2}{n} (\log n)^{5d/4}$$

if $s = 1$, where C_d is a positive constant depending on d.

Remark 4. *Consider the case where $f^* \in \mathcal{F}_{tc}^{d,s}$. By Proposition 5, we then have $f^* \in \mathcal{F}_{tcp}^{d}$, and f^* satisfies* (4.16) *for every $S \subseteq [d]$ with $|S| > s$, so that the assumptions of Theorem 9 are satisfied. Moreover, it follows directly from the definition of $V_{design}(\cdot)$ that*

$$V_{design}(f^*) \leq V_{mars}(f^*).$$

Hence, in this case, the bounds of Theorem 9 hold with $V = V_{mars}(f^*)$, which is independent of n and of the design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$.

By Theorem 9, whenever $f^* \in \mathcal{F}_{\text{tcp}}^{d}$ satisfies (4.16) for every $S \subseteq [d]$ with $|S| > s$, and $V_{\text{design}}(f^*)$ is bounded independently of n (in particular, when $f^* \in \mathcal{F}_{\text{tc}}^{d,s}$), our estimator $\hat{f}_n^{d,s}$ attains the nearly dimension-free rate $n^{-4/5}$, up to logarithmic factors. The rate of the least squares estimator in univariate concave regression is also of order $n^{-4/5}$ ([66, 27, 70, 26, 7]). Theorem 9 therefore shows that our estimator achieves essentially the same rate, differing only by logarithmic factors, even when $s = d$. Thus, totally concave regression largely avoids the curse of dimensionality, making it attractive for regression in higher dimensions.

Shape-constrained estimators often exhibit adaptivity by achieving nearly parametric convergence rates at specific points of the function space (see, e.g., [128, 67]). For $\hat{f}_n^{d,s}$, such adaptivity arises when $f^* \in \mathcal{F}_{\text{tc}}^{d,s}$ also belongs to $\mathcal{R}^{d,s}$, the class of continuous rectangular piecewise multi-affine functions of interaction order at most s. A function f belongs to $\mathcal{R}^{d,s}$ if it is continuous and there exists an axis-aligned split (recall (2.1)) of $[0,1]^d$ of the form $\prod_{j=1}^{d}[v_{l_j-1}^{(j)}, v_{l_j}^{(j)}]$ where $0 = v_0^{(j)} < \cdots < v_{m_j}^{(j)} = 1$ for $j \in [d]$, such that on each rectangle,

$$f(x_1, \dots, x_d) = c_0 + \sum_{0<|S|\leq s} c_S \prod_{j \in S} x_j \tag{4.23}$$

for some constants c_0 and c_S. For $f \in \mathcal{R}^{d,s}$, we denote by $m(f)$ the minimum number of rectangles needed for such a representation. In the univariate case ($d = s = 1$), $\mathcal{R}^{d,s}$ simply consists of continuous piecewise affine functions.

The next theorem shows that if $f^* \in \mathcal{F}_{\text{tc}}^{d,s} \cap \mathcal{R}^{d,s}$, then $\hat{f}_n^{d,s}$ achieves the rate of order n^{-1} (with logarithmic factors), which is much faster than the rate $n^{-4/5}$ given by Theorem 9.

Theorem 10. *Assume the equally-spaced lattice design* (4.22) *and that $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Then, for every $f^* \in \mathcal{F}_{tc}^{d,s} \cap \mathcal{R}^{d,s}$,*

$$R_F(\hat{f}_n^{d,s}, f^*) \leq \begin{cases} C_d \cdot m(f^*) \cdot \sigma^2 (\log n)^{(5d+6s-3)/4}/n & \text{if } s \geq 2 \\ C_d \cdot m(f^*) \cdot \sigma^2 (\log n)^{\max(d+2, 5d/4)}/n & \text{if } s = 1, \end{cases}$$

where C_d is a positive constant depending on d.

The next result generalizes Theorem 10 to cases where f^* need not be rectangular piecewise multi-affine, but may be well-approximated by that class. It establishes a sharp oracle inequality (in the sense of [7]) and applies even when f^* is not totally concave.

Theorem 11. *Assume the equally-spaced lattice design (4.22) and that $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. For every function $f^* : [0,1]^d \to \mathbb{R}$,*

$$R_F(\hat{f}_n^{d,s}, f^*) \le \inf_{f \in \mathcal{F}_{tc}^{d,s} \cap \mathcal{R}^{d,s}} \left\{ \|f - f^*\|_n^2 + C_d \cdot m(f) \cdot \frac{\sigma^2}{n} (\log n)^{(5d+6s-3)/4} \right\} \quad \text{if } s \ge 2,$$

and

$$R_F(\hat{f}_n^{d,1}, f^*) \le \inf_{f \in \mathcal{F}_{tc}^{d,1} \cap \mathcal{R}^{d,1}} \left\{ \|f - f^*\|_n^2 + C_d \cdot m(f) \cdot \frac{\sigma^2}{n} (\log n)^{\max(d+2, 5d/4)} \right\} \quad \text{if } s = 1,$$

where C_d is a positive constant depending on d.

Random Design

In this subsection, we assume that $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ are i.i.d. random vectors with density p_0 on $[0,1]^d$, where p_0 is bounded above by some positive constant B, i.e.,

$$0 \le p_0(\mathbf{x}) \le B \quad \text{for all } \mathbf{x} \in [0,1]^d. \tag{4.24}$$

We also assume that $y_1, \dots, y_n$ are generated according to the model (4.21), where the error terms ξ_i are i.i.d. with mean zero and are independent of $\mathbf{x}^{(i)}$. For an estimator $\hat{f}_n$ of f^*, we measure its accuracy by

$$\|\hat{f}_n - f^*\|_{p_0,2}^2 := \int \left(\hat{f}_n(\mathbf{x}) - f^*(\mathbf{x})\right)^2 p_0(\mathbf{x})\, d\mathbf{x}.$$

Here, we analyze the constrained variant $\hat{f}_{n,V}^{d,s}$ (defined in (4.18)) rather than $\hat{f}_n^{d,s}$. At present, we do not know how to directly analyze the theoretical accuracy of $\hat{f}_n^{d,s}$. It should be noted, however, that in random-design settings, risk upper bounds are not known even for univariate concave regression, which corresponds to $\hat{f}_n^{d,s}$ when $d = s = 1$. Our focus on $\hat{f}_{n,V}^{d,s}$ is also consistent with the existing literature (e.g., [70, 102, 88]), where random-design analyses are carried out for constrained variants of the original least squares estimators. In particular, [70] studies the constrained variant

$$\hat{f}^1_{\text{concave},M} \in \underset{\substack{f \text{ is concave} \\ \|f\|_\infty \le M}}{\operatorname{argmin}} \sum_{i=1}^n \left(y_i - f(x^{(i)})\right)^2,$$

instead of $\hat{f}^1_{\text{concave}}$ (defined in (4.2)), which is the least squares estimator over all univariate concave functions that are uniformly bounded by some constant $M > 0$.

For our first result, we further assume that the error terms ξ_i have finite $L^{5,1}$ norm, i.e.,

$$\|\xi_i\|_{5,1} := \int_0^\infty (\mathbb{P}(|\xi_i| > t))^{1/5}\, dt < \infty. \tag{4.25}$$

This assumption is much weaker than the Gaussian error assumption for our fixed-design results. The additional regularization in $\hat{f}^{d,s}_{n,V}$ allows us to work under this weaker condition.

The following result is a random-design analogue of Theorem 9 and leads to the same qualitative conclusions. It is essentially a corollary of Theorem 5 for MARS. Recall that $\hat{f}^{d,s}_{n,V}$ is the least squares estimator over the class in (4.18), which is contained in the class over which the MARS estimator is optimized (recall (3.5)). Moreover, the assumptions imposed here are the same as those used in the random-design analysis of MARS (see Section 3.3).

Theorem 12. *Suppose $f^* \in \mathcal{F}^{d,s}_{tc}$ and $V_{mars}(f^*) \leq V$. Assume that the density p_0 satisfies* (4.24) *and that the error terms ξ_i satisfy* (4.25)*. Then,*

$$\|\hat{f}^{d,s}_{n,V} - f^*\|^2_{p_0,2} = O_p(n^{-4/5}(\log n)^{8(s-1)/5}),$$

where the constants underlying $O_p(\cdot)$ depend on d, V, B, and the moments of ξ_i.

Our next result shows adaptivity of $\hat{f}^{d,s}_{n,V}$ when $f^* \in \mathcal{F}^{d,s}_{\mathrm{tc}} \cap \mathcal{R}^{d,s}$. For this result, rather than assuming (4.25), we assume that ξ_i are sub-exponential errors satisfying

$$2K^2 \cdot \mathbb{E}\Big[\exp\Big(\frac{|\xi_i|}{K}\Big) - 1 - \frac{|\xi_i|}{K}\Big] \leq \sigma^2 \quad \text{for some } K, \sigma > 0. \tag{4.26}$$

For the density p_0, in addition to (4.24), we further assume that it is bounded below by some positive constant b, i.e.,

$$p_0(\mathbf{x}) \geq b \quad \text{for all } \mathbf{x} \in [0,1]^d. \tag{4.27}$$

Theorem 13. *Suppose $f^* \in \mathcal{F}^{d,s}_{tc} \cap \mathcal{R}^{d,s}$ and $V_{mars}(f^*) \leq V$. Assume that the density p_0 satisfies* (4.24) *and* (4.27) *and that the error terms ξ_i satisfy* (4.26)*. Then,*

$$\|\hat{f}^{d,s}_{n,V} - f^*\|^2_{p_0,2} = O_p\big(n^{-1/2}(\log n)^{\max(5d/8+(s-1)/e,\, d/2+1/2e)}\big),$$

where the constants underlying $O_p(\cdot)$ depend on d, V, b, B, K, σ, and $m(f^*)$.

In all our results, under both fixed and random designs, the covariate domain is $[0,1]^d$. By a scaling argument, the results extend to any axis-aligned rectangle. However, extensions to more general domains (e.g., polytopes) do not follow from our proofs, which rely on rectangular geometry—decompositions into subrectangles and entropy bounds for totally concave function classes on each subrectangle. It is unclear how to adapt our arguments beyond rectangular domains. We also note that total concavity, like axial and additive concavity, is inherently axis-aligned, unlike general concavity. While results for least squares under general concavity exist on non-rectangular domains (see, e.g., [70, 88, 89]), establishing analogous results for the axis-aligned notion of total concavity appears challenging.

4.4 Real Data Examples

We apply our methods to three real datasets and compare them to existing approaches. Implementations of our methods are available in the R package `regmdc` (`https://github.com/DohyeongKi/regmdc`), which also includes our lasso-based variant of MARS in Chapter 3. We first normalize each covariate to $[0, 1]$, and then, after computing our estimators, we invert the transformation and present results on the original scale. The code for all analyses in this section is available at `https://github.com/DohyeongKi/tc-reg-paper`.

In regression, one typically begins with linear models and then adds non-linear terms in each covariate. Among various non-linear terms, quadratic terms are one of the simplest and most common choices. A function that is quadratic in a variable x_j is effectively convex or concave in x_j. Additive models with convex/concave components thus represent nonparametric generalizations of quadratic regression. On the other hand, adding simple product-form interaction terms to a quadratic model lets us move beyond the additive framework.

Totally concave regression (and its variants) provides a natural, nonparametric, shape-constrained alternative that encompasses these additive models and parametric models with interactions, yet remains simple and more interpretable than black-box approaches (e.g., random forests, deep neural networks). Also, in contrast to black-box approaches, which often require extensive tuning, totally concave regression is fully shape-constrained and tuning-free. Our real data examples show that totally concave regression performs comparably to parametric models and improves further when regularization is introduced.

Earnings Data

Consider the classical regression problem (with $d = 2$) from labor economics for which y is the log of weekly earnings, x_1 is years of education, and x_2 is years of experience, for adults in the workforce. Labor economics literature suggests that the regression function should be axially concave (see, e.g., [92, Section III]). On the other hand, classical regression models for this problem (e.g., [107, 108]) are additive and thus inevitably miss crucial interaction effects (see, e.g., [92, Section V]). Totally concave regression presents a natural approach here as it is capable of capturing interactions while staying in the broad framework of axial concavity.

We use the dataset ex1029 from the R library Sleuth3, which contains data on full-time male workers in 1987. We pre-process the data (following [146]) by restricting to non-black workers with $x_1 \geq 8$ and $x_2 \in (0, 40)$ and work with $n = 20,967$ observations. We fit the following models and compare their performance:

- Mincer's [107]: $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_2^2$, where $\mathbf{x} = (x_1, x_2)$.
- Murphy and Welch's [108] (MW): $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_2^2 + c_4 x_2^3 + c_5 x_2^4$.
- Additive Concave Regression: $\mathbb{E}[y|\mathbf{x}] = f_1(x_1) + f_2(x_2)$ where f_1 and f_2 are concave. The least squares estimator over all additive concave functions is $\hat{f}_n^{d,1}$.

- Axially Concave Regression. We fit the least squares estimator over the class of axially concave functions (implementation details are in Appendix B.1).

- Totally Concave Regression. We compute the estimator $\hat{f}_n^{d,s}$ for $d = s = 2$.

Model performance can vary significantly with sample size, making it essential to assess prediction accuracy across different training data sizes. For each $k = 0, \ldots, 8$, we randomly select $0.9/2^k$ of the observations for training and use the remaining $(1 - 0.9/2^k)$ for testing. For instance, $k = 0$ corresponds to a 90–10 train-test split, while $k = 4$ uses 0.9/16 of the observations ($n = 1,179$) for training and the rest for testing. We measure performance using mean squared errors on the test set and repeat each experiment 100 times.

Plot A of Figure 4.2 displays the proportion of repetitions in which each method outperforms the other four (x-axis represents training data proportion); e.g., when the proportion of training data is 0.9/16, the proportions of repetitions in which Mincer, MW, additive, axially, and totally concave regressions perform the best are 0.01, 0.29, 0.06, 0.02, and 0.62. Plot B of Figure 4.2 shows the proportion of repetitions in which each model ranks among the two *worst-performing* models.

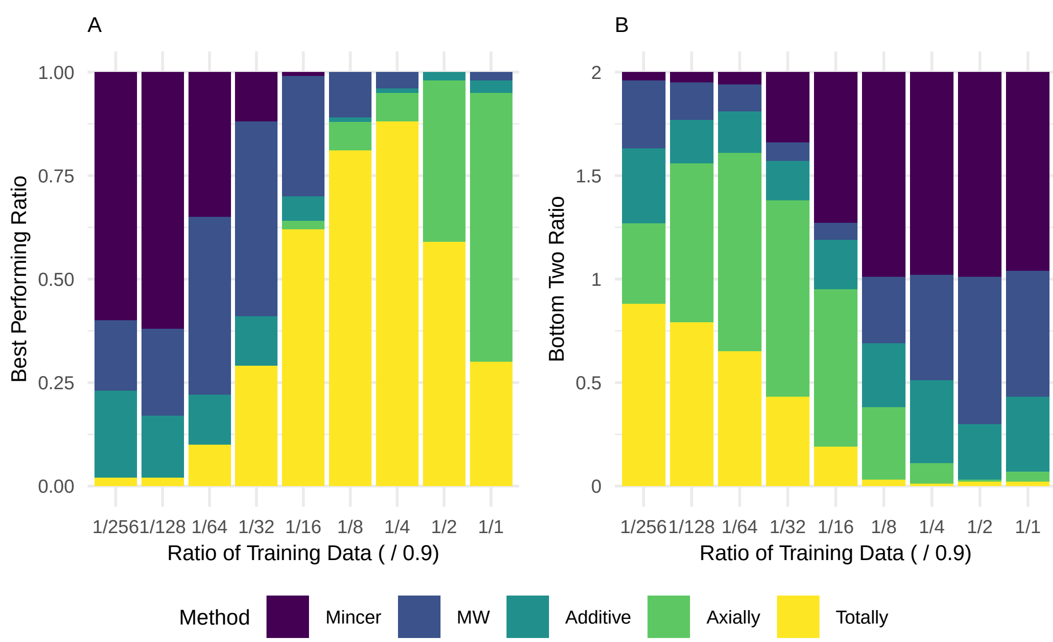

Figure 4.2: Best-performing and bottom-two ratios of each method for various sub-sample sizes. The x-axis and y-axis represent the proportion of training data and the proportion of repetitions (out of 100) where (A) each model outperforms all the other four models and (B) each model is one of the two worst-performing models.

From these plots, it is clear that restrictive models like Mincer's parametric approach perform the best with limited data, but totally concave regression emerges as the preferred model at moderate sample sizes. It maintains this advantage until sample sizes become very large, where axially concave regression takes the lead. This pattern highlights totally concave regression's balanced approach: it captures interactions while maintaining enough structure to perform well with moderately-sized datasets. Although the transition points may vary across datasets, totally concave regression appears to occupy a sweet spot between overly restrictive parametric models and the more flexible axially concave regression.

See Appendix B.2 for additional plots for the analysis.

The above comparison only concerns prediction accuracy and does not account for computational considerations. Axially concave regression becomes computationally prohibitive when $d \geq 3$. The computational burden of totally concave regression can be alleviated by restricting interaction orders (as in $\mathcal{F}_{\text{tc}}^{d,s}$ vs $\mathcal{F}_{\text{tc}}^{d}$), by using approximate versions (as in Section 3.2), or by considering variants (as in Section 4.2). However, it remains unclear whether similar strategies can be adapted to axially concave regression. Due to this computational limitation, axially concave regression will not be considered in the following subsections.

Housing Price Data

We use the hprice2 dataset from the R package wooldridge, which contains housing data from 506 Boston census tracts (1970 US census). The dataset includes median housing price (price), per-capita crime rate (crime), average number of rooms per dwelling (room), nitric oxides level (nox), weighted distance to major employment centers (distance), and student-teacher ratio (stratio). Here are some simple parametric regressions for $y = \log(\textsf{price})$ using $x_1 = \textsf{crime}$, $x_2 = \textsf{room}$, $x_3 = \log(\textsf{nox})$, $x_4 = \log(\textsf{distance})$, and $x_5 = \textsf{stratio}$ as regressors:

- Linear: $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + c_5x_5$, where $\mathbf{x} = (x_1, \dots, x_5)$.
- Quadratic: $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_{11}x_1^2 + c_2x_2 + c_{22}x_2^2 + c_3x_3 + c_4x_4 + c_5x_5$.
- Interaction 1 (Int 1): $\mathbb{E}[y|\mathbf{x}] = c_0+c_1x_1+c_{11}x_1^2+c_2x_2+c_{22}x_2^2+c_{12}x_1x_2+c_3x_3+c_4x_4+c_5x_5$.
- Interaction 2 (Int 2): $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_{11}x_1^2 + c_2x_2 + c_{22}x_2^2 + c_{12}x_1x_2 + c_{112}x_1^2x_2 + c_{122}x_1x_2^2 + c_3x_3 + c_4x_4 + c_5x_5$.
- Interaction 3 (Int 3): $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_{11}x_1^2 + c_2x_2 + c_{22}x_2^2 + c_{12}x_1x_2 + c_3x_3 + c_{13}x_1x_3 + c_{113}x_1^2x_3 + c_{23}x_2x_3 + c_{223}x_2^2x_3 + c_4x_4 + c_5x_5$.
- Interaction 4 (Int 4): $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_{11}x_1^2 + c_2x_2 + c_{22}x_2^2 + c_{12}x_1x_2 + c_{112}x_1^2x_2 + c_{122}x_1x_2^2 + c_3x_3 + c_{13}x_1x_3 + c_{113}x_1^2x_3 + c_{23}x_2x_3 + c_{223}x_2^2x_3 + c_4x_4 + c_5x_5$.

We compare the above models to nonparametric approaches based on total concavity. The quadratic model gives convex (as opposed to concave) fits, so we work with convex variants of our function classes (these are given by functions of the form (4.1) where the negative sign on the last term is replaced by the positive sign). We indicate this by the subscript *convex*.

- Additive ($\hat{f}_n^{5,2,1}$): $\mathbb{E}[y|\mathbf{x}] \in \mathcal{F}^{2,1}_{\text{tc, convex}}(x_1, x_2) \oplus \mathcal{L}(x_3, x_4, x_5)$.
- Our Model 1 ($\hat{f}_n^{5,2,2}$) (Ours 1): $\mathbb{E}[y|\mathbf{x}] \in \mathcal{F}^{2,2}_{\text{tc, convex}}(x_1, x_2) \oplus \mathcal{L}(x_3, x_4, x_5)$.
- Our Model 2 ($\hat{f}_n^{5,3,2,2}$) (Ours 2): $\mathbb{E}[y|\mathbf{x}] \in \mathcal{F}^{3,2}_{\text{tc, convex}}(x_1, x_2; x_3) \oplus \mathcal{L}(x_4, x_5)$.

We evaluated these models using 100 random training-test splits (90%–10%) and ranked them by test mean squared errors (with rank 1 being the best, and rank 9 the worst). Figure 4.3 shows the empirical cumulative distributions of these rankings across the 100 splits. For plots of this kind (including Figures 4.5 and 4.6 appearing later), an ideal procedure is one whose rank distribution function dominates (is everywhere larger than) all the others.

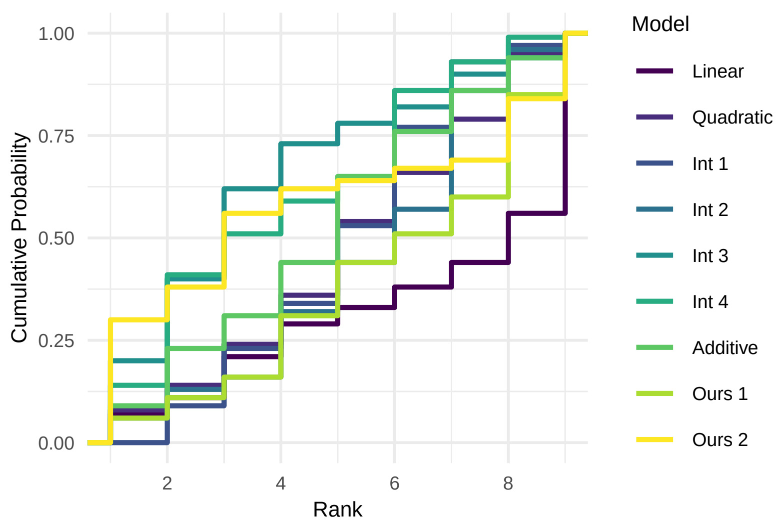

Figure 4.3: Empirical cumulative distributions of model ranks across 100 splits into training and test sets. Ranks are determined by mean squared errors on the test sets.

The model Ours 2 showed both promise and limitations—ranking first in 30% of splits but in the bottom two for another 31%. By contrast, Int 3, Int 4, and Additive demonstrated better consistency, rarely performing poorly though achieving fewer top rankings as well. The remaining models showed no notable advantages.

Figure 4.4 illustrates why Ours 2 performed poorly sometimes. Comparing fitted functions from Ours 2 and Int 3 for a split where Ours 2 ranked worst reveals severe overfitting at one corner of the covariate domain. The dramatically different scales of log(price) between the two panels highlight this problem.

To address overfitting, we introduced regularized variants of Ours 2 ($\hat{f}_{n,V}^{5,3,2,2}$) and Additive ($\hat{f}_{n,V}^{5,2,1}$) (the regularization parameter V is tuned by 10-fold cross-validation, separately for

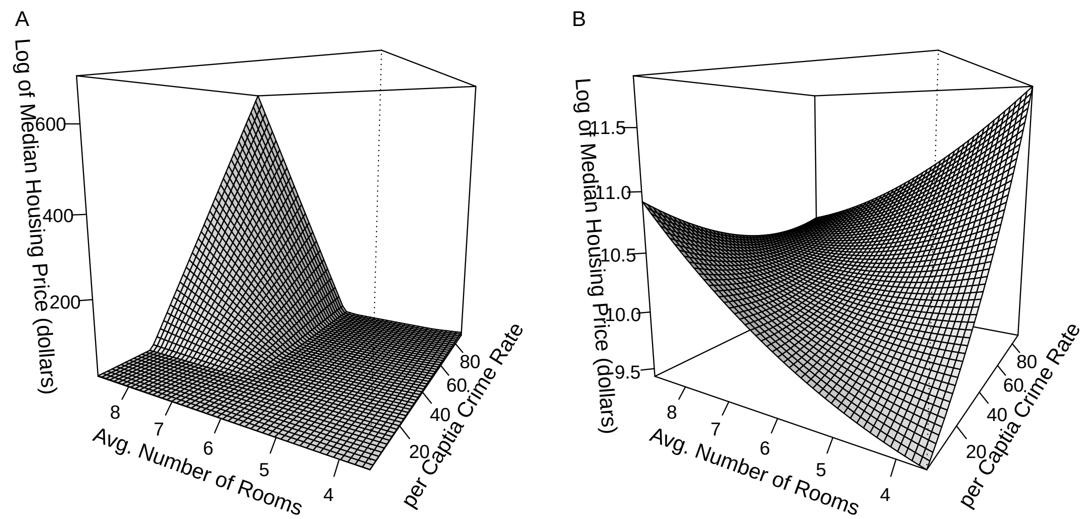

Figure 4.4: Fitted regression functions of (A) Ours 2 and (B) Int 3 for a random split where Ours 2 performs the worst. The regression functions are plotted against crime and room, with log(nox), log(distance), and stratio fixed at their median.

each random split). We excluded consistently underperforming models (Int 1, Int 2, and Ours 1) for clarity and redid the comparison with the regularized variants.

Figure 4.5 shows that the regularized variant of Ours 2 clearly outperformed all other models. Its cumulative distribution dominates all competitors, achieving top-two rankings in 62% of splits. The regularization also effectively addressed the overfitting issue—while the original Ours 2 ranked worst in 28% of splits, its regularized version appeared in the bottom four only 12% of the time. This analysis demonstrates that variants of totally concave regression can outperform standard regression approaches in practical applications.

Retirement Saving Plan (401(k)) Data

We analyzed the 401(k) dataset k401ksubs from the R package wooldridge, containing $n = 9,275$ observations. Our target variable was net total financial assets (y), and predictors were annual income (x_1), age (x_2), 401(k) eligibility (x_3; 1 if eligible, 0 otherwise), and family size. We converted the family size variable (fsize) into five dummy variables: fsize1 (x_4) through fsize4 (x_8) (each equal to 1 if family size equals the corresponding number, 0 otherwise) and fsize5 (equal to 1 if family size ≥ 5). We compare the following models:

- Quadratic: $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_{11}x_1^2 + c_2x_2 + c_{22}x_2^2 + c_3x_3 + c_4x_4 + \cdots + c_8x_8$.
- Int 1: $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_{11}x_1^2 + c_2x_2 + c_{22}x_2^2 + c_{12}x_1x_2 + c_3x_3 + c_4x_4 + \cdots + c_8x_8$.

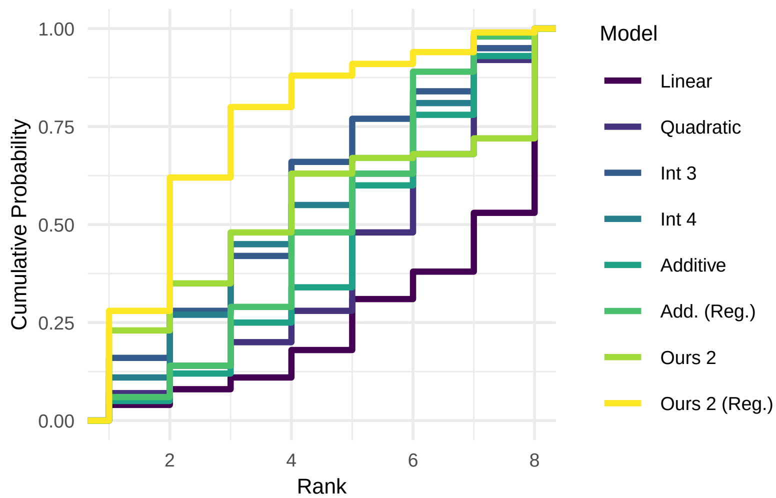

Figure 4.5: Empirical cumulative model rank distributions across 100 training-test splits. Add. (Reg.) and Ours 2 (Reg.) are the regularized variants (4.18) of Additive and Ours 2.

- Int 2: $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_{11}x_1^2 + c_2x_2 + c_{22}x_2^2 + c_{12}x_1x_2 + c_{112}x_1^2x_2 + c_{122}x_1x_2^2 + c_3x_3 + c_4x_4 + \cdots + c_8x_8$.
- Int 3: $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_{11}x_1^2 + c_2x_2 + c_{22}x_2^2 + c_{12}x_1x_2 + c_3x_3 + c_{13}x_1x_3 + c_{113}x_1^2x_3 + c_{23}x_2x_3 + c_{223}x_2^2x_3 + c_4x_4 + \cdots + c_8x_8$.
- Int 4: $\mathbb{E}[y|\mathbf{x}] = c_0 + c_1x_1 + c_{11}x_1^2 + c_2x_2 + c_{22}x_2^2 + c_{12}x_1x_2 + c_{112}x_1^2x_2 + c_{122}x_1x_2^2 + c_3x_3 + c_{13}x_1x_3 + c_{113}x_1^2x_3 + c_{23}x_2x_3 + c_{223}x_2^2x_3 + c_4x_4 + \cdots + c_8x_8$.
- Additive ($\hat{f}_n^{8,2,1}$): $\mathbb{E}[y|\mathbf{x}] \in \mathcal{F}^{2,1}_{\text{tc, convex}}(x_1, x_2) \oplus \mathcal{L}(x_3, \ldots, x_8)$.
- Ours 1 ($\hat{f}_n^{8,2,2}$): $\mathbb{E}[y|\mathbf{x}] \in \mathcal{F}^{2,2}_{\text{tc, convex}}(x_1, x_2) \oplus \mathcal{L}(x_3, \ldots, x_8)$.
- Ours 2 ($\hat{f}_n^{8,3,2,2}$): $\mathbb{E}[y|\mathbf{x}] \in \mathcal{F}^{3,2}_{\text{tc, convex}}(x_1, x_2; x_3) \oplus \mathcal{L}(x_4, \ldots, x_8)$.
- Regularized Variant of Ours 2 ($\hat{f}_{n,V}^{8,3,2,2}$) (Ours 2 (Reg.)).

We focus on total convexity (as opposed to total concavity) as convexity provides better fits (as can be easily verified by inspecting the fits of the simple quadratic model). To manage computational demands, we used approximate versions for Additive, Ours 1, Ours 2, and Ours 2 (Reg.) with proxy lattices with $N_1 = N_2 = 50$ as in (3.18). Since total convexity is only imposed on x_1 and x_2, there is no need for $N_3, \ldots, N_8$.

We evaluated performance across 100 random training-test splits (90%–10%) and ranked models by test mean squared errors. Figure 4.6 shows the empirical rank distributions. The results (from Figure 4.6) align with our findings on housing price data. Ours 2 showed high potential but inconsistency—ranking first in 37 splits but in the bottom two for 29 others. Int 1 and Int 3 demonstrated greater consistency but fewer top placements. The regularized variant of Ours 2 effectively balanced these extremes, ranking worst in only 5 splits while maintaining strong performance. The regularization effect, however, was less dramatic than in the housing price example, likely because the approximation with proxy lattices already provided some overfitting protection by constraining boundary masses. These results confirm totally concave regression's practical utility and demonstrate that approximate versions offer viable alternatives when computational demands are high.

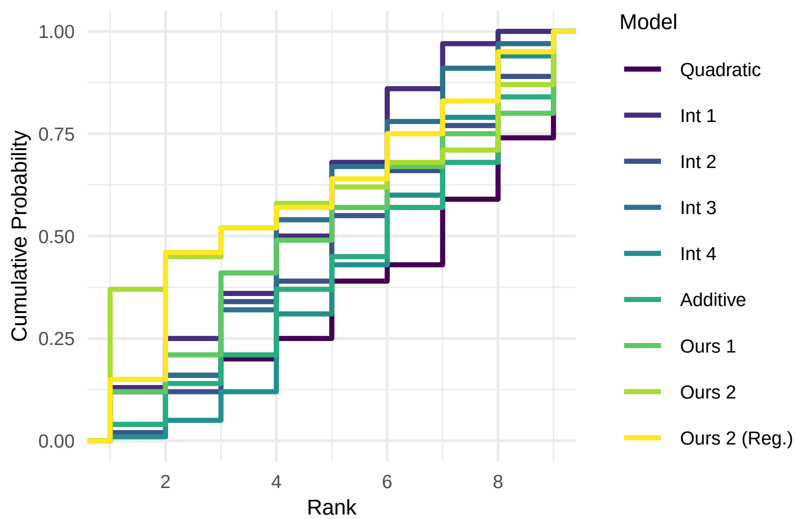

Figure 4.6: Empirical cumulative model rank distributions for 401(k) data

Chapter 5

Extreme Gradient Boosting (XGBoost)

Consider the standard regression problem with data $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$ where each $\mathbf{x}^{(i)} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. XGBoost, introduced by [29], fits a finite sum of regression trees by (approximately) minimizing an objective function consisting of a least squares loss and a regularization penalty. We first describe the optimization problem that XGBoost seeks to solve; see the official documentation [39] for further implementation details.

XGBoost constructs individual regression trees using right-continuous splits, meaning that each split is of the form $x_j \geq t_j$ versus $x_j < t_j$, where x_j denotes the j^{th} coordinate of the covariate vector $\mathbf{x}$. Each tree is further restricted to have a user-specified maximum depth, controlled by the hyperparameter `max_depth` (whose default value is 6). Recall that the depth of a tree refers to the maximum number of splits along any root-to-leaf path.

Let $\mathcal{F}_{\text{st}}^{d,s}$ denote the class of all finite sums of right-continuous[1] regression trees of depth at most s ("st" here stands for "sum of trees"). More precisely, $\mathcal{F}_{\text{st}}^{d,s}$ consists of all functions of the form $\sum_{k=1}^{K} f_k$ for some $K \geq 1$, where each f_k is a regression tree with right-continuous splits and depth $\leq s$. Each f_k can possibly be a constant function, in which case we call it a constant tree and say that it has depth 0. We place no restriction on K beyond finiteness. In implementations, XGBoost allows the user to specify an upper bound on K, typically on the order of several hundred to a few thousand. Since this bound is usually chosen to be large, we leave K unrestricted in the definition of $\mathcal{F}_{\text{st}}^{d,s}$ for theoretical convenience.

For regression, XGBoost minimizes the least squares loss over the class $\mathcal{F}_{\text{st}}^{d,s}$, augmented with an explicit regularization penalty described below. This explicit regularization is a key feature that distinguishes XGBoost from earlier gradient boosting methods such as Gradient Boosting Machines (see [52]), and from ensemble methods such as Random Forests (see [19]).

For a function $f \in \mathcal{F}_{\text{st}}^{d,s}$, suppose f is represented as a finite sum of trees, and let $\mathbf{w}_k$ denote the vector of leaf weights associated with the k^{th} tree. The XGBoost regularization penalty takes the form $\alpha \sum_k \|\mathbf{w}_k\|_1$, where $\alpha > 0$ controls the strength of regularization and

[1] Recall that, in this thesis, the term "right-continuous" means coordinate-wise right-continuity.

$\|\cdot\|_1$ denotes the L^1 norm. If the k^{th} tree is a constant tree, we set $\|\mathbf{w}_k\|_1 = 0$. This penalty discourages overly complex trees by constraining the magnitude of the leaf weights. Since each function $f \in \mathcal{F}_{\text{st}}^{d,s}$ generally admits multiple representations as a finite sum of trees, and since the quantity $\sum_k \|\mathbf{w}_k\|_1$ depends on the particular representation chosen, we obtain a representation-invariant measure of complexity by taking the infimum over all possible tree decompositions. Specifically, for $f \in \mathcal{F}_{\text{st}}^{d,s}$, define

$$V_{\text{xgb}}^{d,s}(f) = \inf\Big\{\sum_k \|\mathbf{w}_k\|_1\Big\}, \tag{5.1}$$

where the infimum is taken over all representations of f as a finite sum of regression trees with right-continuous splits and depth at most s, and $\mathbf{w}_k$ denotes the leaf-weight vector of the k^{th} tree.

In this notation, XGBoost is a greedy algorithm for solving the optimization problem:

$$\operatorname*{argmin}_f \Big\{\sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)})\big)^2 + \alpha V_{\text{xgb}}^{d,s}(f) : f \in \mathcal{F}_{\text{st}}^{d,s}\Big\}. \tag{5.2}$$

It is also common to include an additional leaf-count penalty of the form $\gamma \sum_k T_k$ where T_k is the number of leaves in the k^{th} tree. Since the default value for γ is $\gamma = 0$ (see [39]), we omit this term throughout. Some implementations also replace the L^1 penalty with a squared L^2 penalty $\|\mathbf{w}_k\|_2^2$. However, such a penalty is not well defined for sums of trees. To illustrate this issue, consider the function $(x_1, \dots, x_d) \mapsto \mathbf{1}(x_1 \geq 0)$, for which the squared L^2 penalty is 1. The same function can alternatively be expressed as $\sum_{k=1}^K (1/K) \cdot \mathbf{1}(x_1 \geq 0)$, for which the total squared L^2 penalty equals $1/K$, which tends to zero as $K \to \infty$. The fact that the squared L^2 penalty can be made arbitrarily close to zero by suitably increasing the number of trees in the decomposition—unlike the L^1 penalty—renders the squared L^2 penalty ill-posed for tree ensembles. For this reason, we focus exclusively on the L^1 penalty (5.1). More broadly, the advantages of L^1 over L^2 penalties are well established in high-dimensional statistics (see, e.g., [41, 144, 76, 145]). Additional discussion on the differences between L^1 and L^2 penalties for tree ensembles is provided in Section 5.4.

XGBoost has become one of the most widely used machine learning algorithms, celebrated for its predictive accuracy and efficiency. Indeed, it has played a decisive role in many high-profile machine learning competitions, and practitioners often note that for tabular data (which includes our regression data setting), XGBoost can outperform deep learning methods (see, e.g., [134, 63, 14]). Yet, despite its empirical success, the theoretical properties of XGBoost remain poorly understood, since its iterative boosting procedure, coupled with the greedy tree-building algorithm performed at each iteration, is notoriously difficult to analyze rigorously.

This chapter takes a step toward closing this gap by studying XGBoost through the infinite-dimensional lasso framework (1.5). This framework allows us to identify the function class implicitly underlying XGBoost. It also enables us to extend the complexity measure

(5.1) naturally to a broader context, which helps us characterize and interpret the regularization penalty used by XGBoost. Taken together, these results provide a better understanding of the optimization problem that XGBoost seeks to solve.

Since XGBoost constructs estimators as finite sums of right-continuous regression trees, one natural choice of atoms is the class of such trees. However, instead of applying the infinite-dimensional lasso framework directly to this class, we work with finer building blocks than whole regression trees.

To this end, we begin by observing that any regression tree with right-continuous splits and depth at most s, and hence any finite sum of such trees, can be expressed as a finite linear combination of functions of the form

$$A^{L,U}_{\mathbf{l},\mathbf{u}}(x_1, \dots, x_d) = \prod_{j \in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j \in U} \mathbf{1}(x_j < u_j), \tag{5.3}$$

where L and U are (not necessarily disjoint) subsets of $[d] := \{1, \dots, d\}$ satisfying $|L|+|U| \leq s$ (with $|\cdot|$ denoting set cardinality), and where $\mathbf{l} := (l_j, j \in L)$ and $\mathbf{u} := (u_j, j \in U)$ are vectors of real-valued thresholds. This follows from the fact that each regression tree can be decomposed into indicator functions corresponding to the individual paths from the root to the leaves. For each root-to-leaf path, take

$$\begin{aligned} L &:= \{j \in [d] : \text{the path contains at least one split of the form } x_j \geq t\}, \\ U &:= \{j \in [d] : \text{the path contains at least one split of the form } x_j < t\}, \end{aligned}$$

and for each $j \in L$ (respectively $j \in U$), take l_j (respectively u_j) to be the maximum (respectively minimum) of the thresholds t appearing in such splits along the path. Note that a coordinate may belong to both L and U if it is split in both directions along the same path. Thus, $|L| + |U|$ is bounded above by the depth of the tree, which is at most s.

This shows that the collection of functions of the form (5.3) provides a collection of atoms for XGBoost. Since the thresholds l_j and u_j may vary over all real values, this collection is uncountable. We now apply the infinite-dimensional lasso framework (1.5) to this class of atoms. As in the case of MARS, rather than working with a single signed measure ν as in (1.3), we consider one signed measure $\nu_{L,U}$ for each pair of subsets L and U of $[d]$ satisfying $0 < |L| + |U| \leq s$. This leads to functions of the form

$$f_{c,\{\nu_{L,U}\}}(x_1, \dots, x_d) = c + \sum_{0<|L|+|U| \leq s} \int_{\mathbb{R}^{|L|+|U|}} A^{L,U}_{\mathbf{l},\mathbf{u}}(x_1, \dots, x_d) \, d\nu_{L,U}(\mathbf{l}, \mathbf{u}), \tag{5.4}$$

where $c \in \mathbb{R}$, the summation is over all pairs (L, U) satisfying $0 < |L| + |U| \leq s$, and for each such pair, $\nu_{L,U}$ is a finite signed (Borel) measure on $\mathbb{R}^{|L|+|U|}$. We denote the collection of all such functions by $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$.

As we will see in Section 5.1, $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ contains the original function class $\mathcal{F}^{d,s}_{\mathrm{st}}$ as a subclass, with functions in $\mathcal{F}^{d,s}_{\mathrm{st}}$ corresponding to discrete signed measures $\nu_{L,U}$ having finite support (see Proposition 8). Moreover, unlike $\mathcal{F}^{d,s}_{\mathrm{st}}$, which consists only of piecewise constant functions, $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ contains many continuous functions as well (see Proposition 11).

The lasso-based framework (1.5) also naturally provides (through (1.4)) a measure of complexity for functions in $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$. For $f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$, this complexity is defined by

$$V^{d,s}_{\infty-\mathrm{xgb}}(f) := \inf\left\{ \sum_{0<|L|+|U|\le s} \|\nu_{L,U}\|_{\mathrm{TV}} : f_{c,\{\nu_{L,U}\}} \equiv f \right\}, \tag{5.5}$$

where the infimum is taken over all representations $f_{c,\{\nu_{L,U}\}}$ of f. Here, for a finite signed measure ν, $\|\nu\|_{\mathrm{TV}} := |\nu|(\mathbb{R}^{|L|+|U|})$ denotes its total variation, and the constant term (corresponding to $A^{L,U}_{\mathbf{l},\mathbf{u}}$ with $L = \emptyset = U$) is excluded from the complexity measure. Unlike the case of MARS, where the representation is unique, functions in $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ may admit multiple representations, and hence the infimum is necessary. We will see in Section 5.1 that this complexity measure agrees with the original complexity $V^{d,s}_{\mathrm{xgb}}(\cdot)$ on $\mathcal{F}^{d,s}_{\mathrm{st}}$ (see Theorem 14). In other words, this complexity measure extends $V^{d,s}_{\mathrm{xgb}}(\cdot)$ from $\mathcal{F}^{d,s}_{\mathrm{st}}$ to the larger class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$.

The resulting optimization problem arising from the infinite-dimensional lasso framework can be written as

$$\operatorname*{argmin}_f \left\{ \sum_{i=1}^n \left(y_i - f(\mathbf{x}^{(i)})\right)^2 + \alpha V^{d,s}_{\infty-\mathrm{xgb}}(f) : f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}} \right\}. \tag{5.6}$$

We will also show in Section 5.1 that this optimization problem is equivalent to the original formulation (5.2) (see Theorem 15). More precisely, every solution to (5.2) is also a solution to (5.6), which is posed over the larger class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ and uses the extension $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ of the original measure of complexity.

In Section 5.3, under standard random-design regression assumptions with squared error loss, we also study the minimax rate of convergence over the class

$$\left\{ f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}} : V^{d,s}_{\infty-\mathrm{xgb}}(f) \le V \right\} \tag{5.7}$$

for fixed $V > 0$. We show that the minimax rate of convergence lies between

$$\Omega(n^{-2/3}(\log n)^{2(\min(s,d)-1)/3}) \ \le\ \text{minimax rate} \ \le\ O(n^{-2/3}(\log n)^{4(\min(s,d)-1)/3}), \tag{5.8}$$

where the constants underlying the $\Omega(\cdot)$ and $O(\cdot)$ notations depend on d, s, and V. In particular, both bounds increase with V, indicating that the minimax rate deteriorates as V increases, as one would intuitively expect. The upper bound in (5.8) is achieved by a least squares estimator over (5.7), which can be viewed as solving a constrained version of (5.6). By a standard duality argument, for each $V > 0$, there exists α, possibly depending on both V and the data, such that a solution to the problem (5.6) attains the upper bound in (5.8).

Taken together, these results show that XGBoost can be interpreted as implicitly targeting the larger and more expressive function class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$, even though its fitted solutions are constructed as finite sums of regression trees. The fast convergence rates in (5.8), which do not suffer from the usual curse of dimensionality, suggest that XGBoost can accurately estimate functions $f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ provided that their complexity $V^{d,s}_{\infty-\mathrm{xgb}}(f^*)$ is not too large. This perspective offers, at least in part, a theoretical explanation for the strong empirical performance of XGBoost in practice: although real-world regression functions are rarely piecewise constant, XGBoost can still perform well as long as the underlying function lies in $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ with moderate complexity.

At the same time, our results also point to potential limitations of XGBoost. If the true regression function f^* cannot be well approximated by elements of $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ with controlled complexity $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$, then accurate estimation should not be expected (see Section 5.4 for further discussion). Finally, we emphasize that this chapter does not address algorithmic aspects of XGBoost. Our results characterize the statistical properties of solutions to the objective function that XGBoost aims to optimize, rather than guaranteeing that a specific implementation of the algorithm attains these rates (see Section 5.4 for more details).

We also provide in Section 5.2 a smoothness-based characterization of $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ that does not rely on any explicit connection to trees. Specifically, we show that $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ is closely related to the class of functions with finite Hardy–Krause variation. More precisely, in Proposition 11, we prove that $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ coincides with the class of right-continuous functions having finite Hardy–Krause variation and exhibiting no interactions between covariates of order greater than s, in the precise sense formalized in (5.11). We also show in Section 5.2 that the complexity measure $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ is tightly connected to Hardy–Krause variation: it is bounded above by Hardy–Krause variation, and bounded below by a constant (depending on s and d) times Hardy–Krause variation (see Proposition 12).

Recall from Chapter 2 that Hardy–Krause variation can be interpreted as a measure of smoothness. For sufficiently smooth functions, Hardy–Krause variation is closely related to the L^1 norms of mixed partial derivatives of maximum order one (recall (2.7)). Hence, the above results suggest that XGBoost can be viewed as performing smoothness-constrained nonparametric regression, where smoothness is quantified through control of mixed partial derivatives of maximum order one. Tree-based methods such as XGBoost are often classified as belonging to the "algorithmic modeling" tradition, distinct from statistical modeling (see, e.g., [20]). In contrast, our results place XGBoost squarely within a traditional statistical framework of regularized estimation governed by an interpretable smoothness penalty.

Moreover, recall that Hardy–Krause variation requires the specification of an anchor point. The presence of such an anchor point breaks the symmetry of the domain and thereby introduces an inherent asymmetry into Hardy–Krause variation. Consequently, regression methods based on Hardy–Krause variation depend on the choice of an anchor point, or equivalently, on the orientation of the coordinate axes. In contrast, we will see in Section 5.2 that $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ does not suffer from this lack of symmetry. Owing to this symmetry, $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ provides a more natural regularizer than Hardy–Krause variation, suggesting that XGBoost offers a more natural regression framework than methods based on Hardy–Krause variation.

The results presented in this chapter are taken from [81]. We refer the reader to Appendix E for the proofs of all results stated in this chapter.

5.1 Connection to the Original Formulation

In this section, we first show that the proposed function class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ and complexity measure $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ extend the original function class $\mathcal{F}^{d,s}_{\mathrm{st}}$ and complexity measure $V^{d,s}_{\mathrm{xgb}}(\cdot)$. We also present several basic properties of $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ and $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$. Finally, we study the equivalence between the original optimization problem (5.2) over $\mathcal{F}^{d,s}_{\mathrm{st}}$ with regularization based on $V^{d,s}_{\mathrm{xgb}}(\cdot)$ and the optimization problem (5.6) over $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ with regularization based on $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$.

Function Class

Recall that any regression tree with right-continuous splits and depth $\leq s$ can be expressed as a finite linear combination of the atoms $A^{L,U}_{\mathbf{l},\mathbf{u}}$ with $|L| + |U| \leq s$. Hence, every element of $\mathcal{F}^{d,s}_{\mathrm{st}}$ can be written in the form (5.4) with discrete signed measures $\nu_{L,U}$ having finite support. The next result shows that the converse also holds: every function of the form (5.4) with discrete signed measures $\nu_{L,U}$ having finite support belongs to $\mathcal{F}^{d,s}_{\mathrm{st}}$. The original function class $\mathcal{F}^{d,s}_{\mathrm{st}}$ can therefore be identified with the subclass of $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ generated by discrete signed measures $\nu_{L,U}$ with finite support.

Proposition 8. *The class $\mathcal{F}^{d,s}_{st}$ of all finite sums of right-continuous regression trees of depth at most s can be characterized as*

$$\mathcal{F}^{d,s}_{st} = \left\{ f_{c,\{\nu_{L,U}\}} : \nu_{L,U} \text{ are discrete signed measures with finite support} \right\}.$$

The following result records several basic properties of $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$. In particular, $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ becomes larger as s increases, while it remains unchanged once $s \geq d$.

Proposition 9. *(a) Every function in $\mathcal{F}^{d,s}_{\infty-st}$ is right-continuous.*

(b) For $s_1 \leq s_2$, $\mathcal{F}^{d,s_1}_{\infty-st} \subseteq \mathcal{F}^{d,s_2}_{\infty-st}$.

(c) For every $s \geq d$, $\mathcal{F}^{d,s}_{\infty-st} = \mathcal{F}^{d,d}_{\infty-st}$.

(d) The function class $\mathcal{F}^{d,s}_{\infty-st}$ is convex.

Complexity Measure

Basic properties of the complexity measure $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ are summarized below. Note, in particular, that unlike the function class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$, which does not change once s reaches d, the complexity measure $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ continues to decrease with s even when $s > d$, and eventually stabilizes once $s \geq 2d$.

Proposition 10. *(a) For $s_1 \leq s_2$, $V^{d,s_1}_{\infty-xgb}(f) \geq V^{d,s_2}_{\infty-xgb}(f)$ for all $f \in \mathcal{F}^{d,s_1}_{\infty-st}$.*

(b) For every $s \geq 2d$, $V^{d,s}_{\infty-xgb}(\cdot) \equiv V^{d,2d}_{\infty-xgb}(\cdot)$.

(c) $V^{d,s}_{\infty-xgb}(\cdot)$ *is convex on* $\mathcal{F}^{d,s}_{\infty-st}$*; that is, for all* $f, g \in \mathcal{F}^{d,s}_{\infty-st}$ *and* $\lambda \in [0,1]$,

$$V^{d,s}_{\infty-xgb}((1-\lambda)f + \lambda g) \leq (1-\lambda) \cdot V^{d,s}_{\infty-xgb}(f) + \lambda \cdot V^{d,s}_{\infty-xgb}(g).$$

The next result shows that $V^{d,s}_{\infty-\mathrm{xgb}}(f)$ agrees with the original XGBoost complexity $V^{d,s}_{\mathrm{xgb}}(f)$ (defined in (5.1)) whenever $f \in \mathcal{F}^{d,s}_{\mathrm{st}}$. In other words, $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ is an extension of the original complexity measure $V^{d,s}_{\mathrm{xgb}}(\cdot)$ from $\mathcal{F}^{d,s}_{\mathrm{st}}$ to the larger class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$.

Theorem 14. *For every* $f \in \mathcal{F}^{d,s}_{st}$, *we have* $V^{d,s}_{\infty-xgb}(f) = V^{d,s}_{xgb}(f)$.

Optimization Problem

Here, we analyze the optimization problems (5.2) and (5.6). Our first result establishes the existence of minimizers for both problems and shows that every solution to (5.2) also solves (5.6). This result implies that XGBoost can be viewed as implicitly optimizing over the larger class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$, which contains smooth functions as well as piecewise constant ones.

Theorem 15. *There exists a solution to* (5.6) *that also solves* (5.2)*. Moreover, every solution to* (5.2) *is also a solution to* (5.6)*.*

In standard XGBoost implementations, split thresholds for regression trees are typically restricted to midpoints between observed covariate values. More precisely, for each coordinate j, split thresholds are chosen from the midpoints between consecutive observed values of the j^{th} covariate. Let $v^{(j)}_1 < \cdots < v^{(j)}_{n_j}$ denote the distinct observed values of the j^{th} covariate x_j, sorted in increasing order. Note that

$$\{v^{(j)}_1, \ldots, v^{(j)}_{n_j}\} = \{x^{(1)}_j, \ldots, x^{(n)}_j\}$$

where $x^{(i)}_j$ denotes the j^{th} coordinate of the i^{th} data point $\mathbf{x}^{(i)}$. Let $\mathcal{F}^{d,s}_{\mathrm{st,mid}}$ denote the subclass of $\mathcal{F}^{d,s}_{\mathrm{st}}$ consisting of finite sums of right-continuous trees with depth at most s, where each individual tree restricts split thresholds on the j^{th} coordinate to the set:

$$\{(v^{(j)}_1 + v^{(j)}_2)/2, \ldots, (v^{(j)}_{n_j-1} + v^{(j)}_{n_j})/2\}.$$

Then, the XGBoost algorithm can also be viewed as a greedy procedure for solving:

$$\underset{f}{\operatorname{argmin}} \Big\{ \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}^{(i)})\big)^2 + \alpha V^{d,s}_{\mathrm{xgb}}(f) : f \in \mathcal{F}^{d,s}_{\mathrm{st,mid}} \Big\}. \tag{5.9}$$

The following result shows that the problem (5.6) is also closely related to (5.9).

Theorem 16. *There exists a solution to* (5.6) *that also solves* (5.9). *Moreover, every solution to* (5.9) *is also a solution to* (5.6).

The above pair of theorems is a direct consequence of the following lemma, which asserts that for every $f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$, there exists a function in $\mathcal{F}^{d,s}_{\mathrm{st,mid}}$ that agrees with f at every data point $\mathbf{x}^{(i)}$ and has no greater complexity.

Lemma 5. *For every $f \in \mathcal{F}^{d,s}_{\infty-st}$, there exists $f_{c,\{\nu_{L,U}\}} \in \mathcal{F}^{d,s}_{st,mid}$ such that*

(a) $\nu_{L,U}$ are discrete signed measures supported on the lattices

$$\prod_{j\in L}\left\{(v_1^{(j)}+v_2^{(j)})/2,\dots,(v_{n_j-1}^{(j)}+v_{n_j}^{(j)})/2\right\}\times\prod_{j\in U}\left\{(v_1^{(j)}+v_2^{(j)})/2,\dots,(v_{n_j-1}^{(j)}+v_{n_j}^{(j)})/2\right\}; \quad (5.10)$$

(b) $f_{c,\{\nu_{L,U}\}}(\mathbf{x}^{(i)}) = f(\mathbf{x}^{(i)})$ for $i = 1,\dots,n$;

(c)

$$V^{d,s}_{\infty-xgb}(f_{c,\{\nu_{L,U}\}}) = \sum_{0<|L|+|U|\le s} \|\nu_{L,U}\|_{TV} \le V^{d,s}_{\infty-xgb}(f).$$

Lemma 5 continues to hold even if the midpoint $(v^{(j)}_{m_j} + v^{(j)}_{m_j+1})/2$ is replaced by any other point in the interval $(v^{(j)}_{m_j}, v^{(j)}_{m_j+1})$. We use midpoints because this choice aligns with standard XGBoost implementations. By default, XGBoost uses midpoints when the dataset is small, although it switches to quantile-based splits for larger datasets due to computational limitations (see [39]).

The equality in the first part of condition (c) deserves special attention. Since $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ is defined as an infimum over all admissible integral representations (5.4), in general, only an inequality holds between $V^{d,s}_{\infty-\mathrm{xgb}}(f_{c,\{\nu_{L,U}\}})$ and the sum of the total variations of the signed measures $\nu_{L,U}$. However, for the functions constructed in Lemma 5, equality is attained. This eliminates the need to take an infimum and allows the penalty term to be expressed explicitly as a sum of the total variations of the associated signed measures.

5.2 Connection to Hardy–Krause Variation

In this section, we study how the function class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ and complexity measure $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ relate to Hardy–Krause variation. Recall from Chapter 2 the definition of Hardy–Krause variation and its properties. We first show that $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ can be precisely characterized using Hardy–Krause variation. We also discuss similarities and differences between $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ and Hardy–Krause variation. In particular, we show that $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ is always within a constant factor (depending on d and s) of Hardy–Krause variation, while exhibiting a symmetry that Hardy–Krause variation lacks.

Function Class

The next result shows that $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ consists precisely of all right-continuous functions with finite Hardy–Krause variation that satisfy the following interaction restriction condition:

$$\sum_{\boldsymbol{\delta}\in\{0,1\}^{|S|}} (-1)^{\sum_{j\in S}\delta_j} \cdot f^S_{(a_j, j\in S^c)}\big((1-\delta_j)w_j + \delta_j v_j, j \in S\big) = 0 \quad \text{for all } v_j < w_j, j \in S \tag{5.11}$$

for every subset $S \subseteq [d]$ with $|S| > s$. Recall the notation $f^S_{(a_j, j\in S^c)}$ from (2.6). It can be readily verified that $f^S_{(a_j, j\in S^c)}$ is well defined for every $S \subseteq [d]$ and every $\mathbf{a} \in \{-\infty, +\infty\}^d$ whenever $f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$. As (4.16) does for totally concave regression, this condition excludes interactions between covariates of order greater than s. Using the divided difference notation introduced in Chapter 4, it can also be written informally as

$$\begin{bmatrix} v_j, w_j, & j \in S \\ a_j, & j \notin S \end{bmatrix} ; \; f \Bigg] = 0 \quad \text{for every } v_j < w_j, j \in S$$

for every $S \subseteq [d]$ with $|S| > s$.

Proposition 11. *The following statements are equivalent:*

(a) $f \in \mathcal{F}^{d,s}_{\infty-st}$.

(b) *f is right-continuous, and for some $\mathbf{a} \in \{-\infty, +\infty\}^d$, $HK_{\mathbf{a}}(f) < \infty$ and f satisfies* (5.11) *for all subsets $S \subseteq [d]$ with $|S| > s$.*

(c) *f is right-continuous, and for all $\mathbf{a} \in \{-\infty, +\infty\}^d$, $HK_{\mathbf{a}}(f) < \infty$ and f satisfies* (5.11) *for all subsets $S \subseteq [d]$ with $|S| > s$.*

Remark 5 ($d = 1$)**.** *When $d = 1$, Proposition 11 simplifies as follows. For every $s \geq 1$,*

$$\mathcal{F}^{1,s}_{\infty-st} = \big\{ f : TV(f) < \infty \text{ and } f \text{ is right-continuous} \big\}.$$

Here, $TV(f)$ denotes the usual total variation of f on $\mathbb{R}$, defined by

$$TV(f) = \sup_{a<b} TV(f; [a, b]),$$

where

$$TV(f; [a, b]) = \sup \sum_{k=1}^{m} |f(z_k) - f(z_{k-1})|,$$

with the supremum taken over all $m \geq 1$ and all partitions $a = z_0 < \cdots < z_m = b$ of $[a, b]$.

Proposition 11 confirms that $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ contains many continuous functions, in contrast to the subclass $\mathcal{F}^{d,s}_{\mathrm{st}}$, which consists only of piecewise constant functions. For example, any sufficiently smooth function whose mixed partial derivatives of maximum order one have finite L^1 norms (recall (2.7)) belongs to $\mathcal{F}^{d,d}_{\infty-\mathrm{st}}$.

Complexity Measure

When $d = 1$, we have the following explicit formula for $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ in terms of total variation (recall that Hardy–Krause variation coincides with total variation when $d = 1$). By Proposition 10, $V^{1,s}_{\infty-\mathrm{xgb}}(\cdot) \equiv V^{1,2}_{\infty-\mathrm{xgb}}(\cdot)$ for all $s \geq 2$. For $f \in \mathcal{F}^{1,s}_{\infty-\mathrm{st}}$, we have

$$V^{1,s}_{\infty-\mathrm{xgb}}(f) = \begin{cases} \mathrm{TV}(f) & \text{if } s = 1, \\ \frac{1}{2}(\mathrm{TV}(f) + |\Delta(f)|) & \text{if } s = 2, \end{cases} \tag{5.12}$$

where $\Delta(f) := \lim_{x\to+\infty} f(x) - \lim_{x\to-\infty} f(x)$. Since $|\Delta(f)| \leq \mathrm{TV}(f)$, it follows that

$$\frac{1}{2}\mathrm{TV}(f) \leq V^{1,s}_{\infty-\mathrm{xgb}}(f) \leq \mathrm{TV}(f). \tag{5.13}$$

When $d \geq 2$, it does not seem possible to provide a direct formula for $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ in terms of Hardy–Krause variation, but an inequality analogous to (5.13) still holds, as shown in the next result.

Proposition 12. *For every $f \in \mathcal{F}^{d,s}_{\infty-st}$, we have*

$$\frac{1}{\min(2^s - 1, 2^d)} \cdot \Big(\sup_{\mathbf{a}\in\{-\infty,+\infty\}^d} HK_{\mathbf{a}}(f) \Big) \leq V^{d,s}_{\infty-xgb}(f) \leq \inf_{\mathbf{a}\in\{-\infty,+\infty\}^d} HK_{\mathbf{a}}(f). \tag{5.14}$$

Both sides of the inequality are tight, in the sense that there exist non-constant functions in $\mathcal{F}^{d,s}_{\infty-st}$ for which the left and right inequalities hold with equality, respectively.

An important distinction between $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ and Hardy–Krause variation is that $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ is symmetric, whereas Hardy–Krause is inherently asymmetric. For example, when $d = s = 2$ and $\mathbf{a} = (-\infty, -\infty)$, for any $t_1, t_2 \in \mathbb{R}$, we have

$$\mathrm{HK}_{\mathbf{a}}\big((x_1, x_2) \mapsto \mathbf{1}(x_1 \geq t_1, x_2 \geq t_2)\big) = 1,$$

while

$$\mathrm{HK}_{\mathbf{a}}\big((x_1, x_2) \mapsto \mathbf{1}(x_1 < t_1, x_2 \geq t_2)\big) = 2 \quad \text{and} \quad \mathrm{HK}_{\mathbf{a}}\big((x_1, x_2) \mapsto \mathbf{1}(x_1 < t_1, x_2 < t_2)\big) = 3.$$

Similar asymmetry arises for other choices of $\mathbf{a}$. In contrast, for $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$, we have

$$\begin{aligned} &V^{d,s}_{\infty-\mathrm{xgb}}\big((x_1, x_2) \mapsto \mathbf{1}(x_1 \geq t_1, x_2 \geq t_2)\big) = V^{d,s}_{\infty-\mathrm{xgb}}\big((x_1, x_2) \mapsto \mathbf{1}(x_1 < t_1, x_2 \geq t_2)\big) \\ &\quad = V^{d,s}_{\infty-\mathrm{xgb}}\big((x_1, x_2) \mapsto \mathbf{1}(x_1 \geq t_1, x_2 < t_2)\big) = V^{d,s}_{\infty-\mathrm{xgb}}\big((x_1, x_2) \mapsto \mathbf{1}(x_1 < t_1, x_2 < t_2)\big) = 1. \end{aligned}$$

This asymmetry in Hardy–Krause variation arises from the presence of an anchor point. Hardy–Krause variation anchors the function at a single corner of the domain, thereby inducing asymmetry. Consequently, estimation results based on Hardy–Krause variation as a regularization penalty may change if the anchor point is moved to another corner or, equivalently, if some coordinate axes of the domain are flipped. By contrast, $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ is invariant to axis flipping, as formalized in the next proposition, which suggests that $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ provides a more natural regularizer than Hardy–Krause variation $\mathrm{HK}_{\mathbf{a}}(\cdot)$.

Proposition 13. *Let $f \in \mathcal{F}^{d,s}_{\infty-st}$. Fix $j \in [d]$ and $t_j \in \mathbb{R}$, and define $g : \mathbb{R}^d \to \mathbb{R}$ by*

$$g(x_1, \dots, x_d) = f(x_1, \dots, x_{j-1}, t_j - x_j, x_{j+1}, \dots, x_d) \quad \textit{for } (x_1, \dots, x_d) \in \mathbb{R}^d.$$

Then, $V^{d,s}_{\infty-xgb}(g) = V^{d,s}_{\infty-xgb}(f)$.

Additional insight into the asymmetry of Hardy–Krause variation, contrasted with the symmetry of $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$, is provided in Section 5.4.

5.3 Theoretical Accuracy

In this section, we study the minimax rate of convergence over the function class (5.7). Throughout, we assume $(\mathbf{x}^{(1)}, y_1), \dots, (\mathbf{x}^{(n)}, y_n)$ are generated according to the model

$$y_i = f^*(\mathbf{x}^{(i)}) + \xi_i$$

where f^* is the true regression function. We work in the random-design setting, in which the covariates $\mathbf{x}^{(i)}$ are assumed to be i.i.d. with density p_0 supported on a compact rectangle and bounded from above:

$$p_0(\mathbf{x}) = 0 \ \text{ when } \mathbf{x} \notin \prod_{j=1}^{d} \Big[-\frac{M_j}{2}, \frac{M_j}{2} \Big] \quad \text{and} \quad B := M_1 \cdots M_d \cdot \sup_{\mathbf{x}} p_0(\mathbf{x}) < \infty. \tag{5.15}$$

Note that when p_0 is the density of the uniform distribution on $\prod_{j=1}^d [-M_j/2, M_j/2]$, we have $B = 1$. We further assume that the error terms ξ_i are i.i.d., mean-zero, and independent of the covariates $\mathbf{x}^{(i)}$.

The minimax risk over the class (5.7) is defined as

$$\mathfrak{M}^{d,s}_{n,V} := \inf_{\hat{f}_n} \sup_{\substack{f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}} \\ V^{d,s}_{\infty-\mathrm{xgb}}(f^*) \le V}} \mathbb{E} \| \hat{f}_n - f^* \|^2_{p_0,2}, \tag{5.16}$$

where the infimum is taken over all estimators $\hat{f}_n$ based on the data $(\mathbf{x}^{(1)}, y_1), \dots, (\mathbf{x}^{(n)}, y_n)$. Here, $\|\hat{f}_n - f^*\|_{p_0,2}$ denotes the $L^2(p_0)$ loss between $\hat{f}_n$ and f^*:

$$\|\hat{f}_n - f^*\|^2_{p_0,2} := \int_{\mathbb{R}^d} \big(\hat{f}_n(\mathbf{x}) - f^*(\mathbf{x}) \big)^2 p_0(\mathbf{x}) \, d\mathbf{x}.$$

The first main result of this section establishes an upper bound on the minimax risk (5.16). We obtain this bound by analyzing a specific least squares estimator over the class (5.7). Specifically, we consider the least squares estimator over (5.7) subject to the additional restrictions that the associated signed measures $\nu_{L,U}$ satisfy condition (a) and attain equality in condition (c) of Lemma 5:

$$\hat{f}^{d,s}_{n,V} \in \underset{f}{\operatorname{argmin}} \Bigg\{ \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}^{(i)})\big)^2 : f \equiv f_{c,\{\nu_{L,U}\}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}, \tag{5.17}$$
$$\sum_{0<|L|+|U|\le s} \|\nu_{L,U}\|_{\mathrm{TV}} \le V, \text{ and } \nu_{L,U} \text{ satisfy condition (a) of Lemma 5} \Bigg\}.$$

Lemma 5 guarantees that $\hat{f}^{d,s}_{n,V}$ also minimizes the least squares criterion over the original class (5.7). In other words, it is a least squares estimator over the class (5.7):

$$\hat{f}^{d,s}_{n,V} \in \underset{f}{\operatorname{argmin}} \Bigg\{ \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}^{(i)})\big)^2 : f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}} \text{ and } V^{d,s}_{\infty-\mathrm{xgb}}(f) \le V \Bigg\}.$$

One can further verify that for each V, there exists α, possibly depending on V and the data, such that $\hat{f}^{d,s}_{n,V}$ is also a solution to the original penalized formulation (5.6). More precisely, if α is chosen as the solution to the Lagrange dual problem of (5.17), then $\hat{f}^{d,s}_{n,V}$ is also a solution to the penalized version of (5.17) and hence a solution to (5.6) (recall Lemma 5).

The following theorem provides an upper bound on the risk of $\hat{f}^{d,s}_{n,V}$. For this result, we impose an additional assumption on the error terms ξ_i. Specifically, we assume that they have finite $L^{3,1}$ norm:

$$\|\xi_i\|_{3,1} := \int_0^\infty (\mathbb{P}(|\xi_i| > t))^{1/3}\, dt < \infty. \tag{5.18}$$

Recall from Sections 3.3 and 4.3 that, for the random-design analyses of MARS and totally concave regression, we assumed that the error terms ξ_i have finite $L^{5,1}$ norm. Here, we require a weaker assumption. In particular, the finite $L^{3,1}$ norm condition lies between the finite L^3 norm condition and the finite $L^{3+\epsilon}$ norm condition for any $\epsilon > 0$.

Theorem 17. *Fix a true regression function $f^* : \mathbb{R}^d \to \mathbb{R}$, not necessarily belonging to $\mathcal{F}^{d,s}_{\infty-st}$. Suppose that the density p_0 satisfies (5.15) and that the error terms ξ_i satisfy (5.18). Then, for every $f_0 \in \mathcal{F}^{d,s}_{\infty-st}$ with $V^{d,s}_{\infty-xgb}(f_0) < V$, we have*

$$\mathbb{E}\big[\|\hat{f}^{d,s}_{n,V} - f^*\|^2_{p_0,2}\big] \le C\|f_0 - f^*\|^2_{p_0,2} + O\big(d^{4\bar{s}}(1+\log d)^{4(\bar{s}-1)}(V+1)^2 \cdot n^{-2/3}(\log n)^{4(\bar{s}-1)/3}\big), \tag{5.19}$$

where $C > 0$ is a universal constant, $\bar{s} := \min(s,d)$, and the constant factor underlying $O(\cdot)$ depends on B, s, the moments of ξ_i, and

$$\sup_{\mathbf{x}\in\prod_{j=1}^d[-M_j/2, M_j/2]} |f_0(\mathbf{x}) - f^*(\mathbf{x})|.$$

Theorem 17 is stated in a misspecified setting, allowing the true function f^* to be arbitrary. If $f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ and $V^{d,s}_{\infty-\mathrm{xgb}}(f^*) < V$, then we can take $f_0 = f^*$ in (5.19) to deduce

$$\mathbb{E}\big[\|\hat{f}^{d,s}_{n,V} - f^*\|^2_{p_0,2}\big] \le O\big(d^{4\bar{s}}(1+\log d)^{4(\bar{s}-1)}(V+1)^2 \cdot n^{-2/3}(\log n)^{4(\bar{s}-1)/3}\big),$$

where the constant factor underlying $O(\cdot)$ depends on B, s, and the moments of ξ_i. This shows that when $f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$, the least squares estimator $\hat{f}^{d,s}_{n,V}$ over the class (5.7) converges to f^* at the rate $n^{-2/3}$, up to logarithmic factors. The upper bound depends on the complexity bound V on $V^{d,s}_{\infty-\mathrm{xgb}}(f^*)$ through the factor $(V+1)^2$, indicating that the accuracy of $\hat{f}^{d,s}_{n,V}$ deteriorates as the complexity of the target function increases. It is also worth noting that the dependence on d in the bound is polynomial.

Remark 6. *Given the relationship between $V^{d,s}_{\infty-xgb}(\cdot)$ and Hardy–Krause variation discussed in Section 5.2, it is natural to compare Theorem 17 with existing results on Hardy–Krause variation denoising, such as those in [46]. In particular, Theorem 4.5 of [46] shows that the least squares estimator under a Hardy–Krause variation constraint also achieves an $n^{-2/3}$ rate of convergence (up to a slightly different logarithmic factor).*

This similarity is not surprising in light of the close connection between Hardy–Krause variation and $V^{d,s}_{\infty-xgb}(\cdot)$, especially Proposition 12. However, there are important differences. Theorem 4.5 of [46] is established under a fixed-design setting, where the design points $\mathbf{x}^{(i)}$ form a lattice, whereas our result assumes random designs, which are more relevant in many applications. Also, the analysis of [46] is restricted to the case $s = d$ (in their framework, this means all interaction orders between covariates are allowed), while our result holds for all $1 \le s \le d$. Moreover, the bounds in [46] do not explicitly specify the dependence on d.

Remark 7. *It is worth noting that the upper bound in Theorem 17 is deterministic rather than probabilistic, unlike our random-design risk upper bounds for MARS and totally concave regression (recall Theorems 5 and 12). In random-design settings, such deterministic bounds are typically unavailable when the underlying function class is not uniformly bounded. This is because empirical process theory tools that play a central role in random-design analyses often require uniform boundedness of the function class (see, e.g., [150, 58, 69, 84]).*

To overcome this difficulty, one commonly proves that the estimator is bounded in probability, that is, $\|\hat{f}_n\|_\infty = O_p(1)$ (see, e.g., [71, 69, 84, 85]). However, this typically leads to risk bounds that are themselves probabilistic. This is also the reason that our risk bounds for MARS and totally concave regression are probabilistic. In contrast, for XGBoost, although the class $\mathcal{F}^{d,s}_{\infty-st}$ is not uniformly bounded, each function in $\mathcal{F}^{d,s}_{\infty-st}$ can always be decomposed into a constant term plus a uniformly bounded component. We exploit this structural property to obtain deterministic risk bounds despite the lack of uniform boundedness in $\mathcal{F}^{d,s}_{\infty-st}$.

The following upper bound on the bracketing entropy is a key ingredient for the proof of Theorem 17. Let $\mathcal{F}_{\mathbf{M}}(V)$ denote the class of all functions $f_{c,\{\nu_{L,U}\}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ of the form (5.4) satisfying:

- $\nu_{L,U}$ are supported on $\prod_{j\in L}(-M_j/2, M_j/2] \times \prod_{j\in U}(-M_j/2, M_j/2]$;
- $\sum_{0<|L|+|U|\le s} \|\nu_{L,U}\|_{\mathrm{TV}} \le V$.

The class $\mathcal{F}_{\mathbf{M}}(V)$ is not totally bounded, since it contains all constant functions. We therefore restrict attention to the subclass

$$D_{\mathbf{M}}(V,t) = \{f \in \mathcal{F}_{\mathbf{M}}(V) : \|f\|_{p_0,2} \leq t\}.$$

Lemma 6. *There exist constants $C_s > 0$, depending on s, and $C_{B,s} > 0$, depending on B and s, such that for every $\epsilon, t, V > 0$,*

$$\log N_{[\,]}(\epsilon, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2}) \leq \log\Big(2 + \frac{C_s(V+t)}{\epsilon}\Big) + C_{B,s} d^{2\bar{s}}(1+\log d)^{2(\bar{s}-1)}\Big(2 + \frac{V}{\epsilon}\Big)\Big[\log\Big(2+\frac{V}{\epsilon}\Big)\Big]^{2(\bar{s}-1)}.$$

Here, $N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_{p_0,2})$ denotes the ϵ-bracketing number of $\mathcal{F}$ with respect to the norm $\|\cdot\|_{p_0,2}$.

This result builds on the bracketing entropy bounds of [55] for multivariate cumulative distribution functions corresponding to probability measures supported on a fixed bounded axis-aligned rectangle. The connection between the class $\mathcal{F}_{\mathbf{M}}(V)$ and the class of multivariate cumulative distribution functions follows from the observation that each term $\int A^{L,U}_{\mathbf{l},\mathbf{u}}\, d\nu_{L,U}(\mathbf{l},\mathbf{u})$ in (5.4) resembles a cumulative distribution function, since the basis functions $A^{L,U}_{\mathbf{l},\mathbf{u}}$ are constructed from indicator functions.

An earlier work by [13] establishes metric entropy bounds (rather than bracketing entropy bounds) for the same class of cumulative distribution functions, with a sharper logarithmic factor. However, for the proof of Theorem 17, bracketing entropy is essential, and the results of [13] therefore cannot be directly applied.

Theorem 17 immediately implies the following corollary.

Corollary 1. *Suppose that the density p_0 satisfies* (5.15) *and that the error terms ξ_i satisfy* (5.18)*. The minimax risk $\mathfrak{M}^{d,s}_{n,V}$ then satisfies*

$$\mathfrak{M}^{d,s}_{n,V} \leq O\big(d^{4\bar{s}}(1+\log d)^{4(\bar{s}-1)}(V+1)^2 \cdot n^{-2/3}(\log n)^{4(\bar{s}-1)/3}\big),$$

where the constant factor underlying $O(\cdot)$ depends on B, s, and the moments of ξ_i.

We now turn to the second main result of this section, which establishes a lower bound on the minimax risk. For this lower bound result, in addition to (5.15), we further assume that the density p_0 is bounded away from zero on its support, in the sense that

$$b := M_1 \cdots M_d \cdot \inf_{\mathbf{x} \in \prod_{j=1}^d [-M_j/2, M_j/2]} p_0(\mathbf{x}) > 0. \tag{5.20}$$

Note that $b = 1$ when p_0 is the uniform density on $\prod_{j=1}^d [-M_j/2, M_j/2]$. For the error terms ξ_i, instead of (5.18), we assume that they are Gaussian:

$$\xi_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2). \tag{5.21}$$

Theorem 18. *Suppose the density p_0 satisfies* (5.15) *and* (5.20)*, and the error terms ξ_i satisfy* (5.21)*. Then, there exist constants $C_{b,B,\bar{s}} > 0$, depending on b, B, and $\bar{s} = \min(s, d)$, and $C_{B,\bar{s}} > 0$, depending on B and $\bar{s}$, such that*

$$\mathfrak{M}_{n,V}^{d,s} \geq C_{b,B,\bar{s}} \Big(\frac{\sigma^2 V}{n}\Big)^{2/3} \Big[\log\Big(\frac{nV^2}{\sigma^2}\Big)\Big]^{2(\bar{s}-1)/3},$$

provided that $n \geq C_{B,\bar{s}}(\sigma^2/V^2)$.

Combining Corollary 1 and Theorem 18, we conclude that the minimax rate of convergence over the class (5.7) is $n^{-2/3}$, up to logarithmic factors whose exponent lies between $2(\bar{s}-1)/3$ and $4(\bar{s}-1)/3$. This nearly dimension-free rate indicates that the class (5.7) is sufficiently regularized even in high dimensions. In other words, the complexity measure $V_{\infty-\mathrm{xgb}}^{d,s}(\cdot)$ (and hence $V_{\mathrm{xgb}}^{d,s}(\cdot)$) provides effective regularization as the dimension d increases, adequately controlling model complexity in high-dimensional settings.

This observation offers a possible explanation for the strong empirical performance of XGBoost, complementing the fact that every solution to the XGBoost optimization problem (5.2) also solves the penalized least squares problem (5.6) (Theorem 15) over the function class $\mathcal{F}_{\infty-\mathrm{st}}^{d,s}$, which contains many functions beyond piecewise constant ones (Proposition 11).

5.4 Discussion

Connection to a Symmetrized Hardy–Krause Variation

As discussed in Section 5.2, Hardy–Krause variation has a lack of symmetry due to the need to specify an anchor point. One can attempt to restore symmetry by combining all 2^d versions $\mathrm{HK}_{\mathbf{a}}(\cdot)$ corresponding to $\mathbf{a} = (a_1, \dots, a_d)$ with $a_j \in \{-\infty, +\infty\}$. In the mathematical image processing literature, a natural device for combining multiple notions of variation into a single quantity is *infimal convolution* (see, e.g., [25, 132, 133, 9, 17]). Following this idea, one may consider the infimal convolution of the Hardy–Krause variations $\mathrm{HK}_{\mathbf{a}}(\cdot)$ over all $\mathbf{a} \in \{-\infty, +\infty\}^d$:

$$\inf\Big\{\sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} \mathrm{HK}_{\mathbf{a}}(f_{\mathbf{a}}) : \sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} f_{\mathbf{a}} \equiv f,\ f_{\mathbf{a}} \in \mathcal{F}_{\infty-\mathrm{st}}^{d,s}\ \forall \mathbf{a}\Big\}. \tag{5.22}$$

A natural question is then how this quantity, which also satisfies the symmetry condition described in Proposition 13, relates to $V_{\infty-\mathrm{xgb}}^{d,s}(\cdot)$.

It can be shown that if the definitions of $\mathcal{F}_{\infty-\mathrm{st}}^{d,s}$ and $V_{\infty-\mathrm{xgb}}^{d,s}(\cdot)$ are modified to forbid repeated use of the same variable for splits within each tree, then the resulting complexity measure coincides with (5.22). To make this precise, consider the function class $\widetilde{\mathcal{F}}_{\infty-\mathrm{st}}^{d,s}$ consisting of all functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form (5.4), but with the sum ranging only over *disjoint* subsets L and U of $[d]$ satisfying $0 < |L| + |U| \leq s$. For each $f \in \widetilde{\mathcal{F}}_{\infty-\mathrm{st}}^{d,s}$, define the complexity $\widetilde{V}^{d,s}_{\infty-\mathrm{xgb}}(f)$ of f analogously to (5.5), again restricting the sum to disjoint subsets L and U with $0 < |L| + |U| \le s$.

One can verify that $\widetilde{\mathcal{F}}^{d,s}_{\infty-\mathrm{st}} = \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ and that $V^{d,s}_{\infty-\mathrm{xgb}}(f) \le \widetilde{V}^{d,s}_{\infty-\mathrm{xgb}}(f) \le \mathrm{HK}_{\mathbf{a}}(f)$ for all $f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ and all $\mathbf{a} \in \{-\infty, +\infty\}^d$. More importantly, $\widetilde{V}^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ coincides with the infimal convolution in (5.22), as shown in the following proposition.

Proposition 14. *For every $f \in \mathcal{F}^{d,s}_{\infty-st}$, we have*

$$\widetilde{V}^{d,s}_{\infty-xgb}(f) = \inf\Big\{ \sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} HK_{\mathbf{a}}(f_{\mathbf{a}}) : \sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} f_{\mathbf{a}} \equiv f,\ f_{\mathbf{a}} \in \mathcal{F}^{d,s}_{\infty-st}\ \forall \mathbf{a} \Big\}.$$

Although $\widetilde{V}^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ is symmetric and admits a clean characterization via infimal convolution of Hardy–Krause variations across different anchors, it does not fully reflect the behavior of regression trees as used in practice. An important aspect of regression trees is the ability to split on the same variable multiple times within a single tree, which enables localized refinement along a coordinate. Disallowing such repeated splits can reduce estimation accuracy in practice. The complexity $\widetilde{V}^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ corresponds to this restricted setting, whereas $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ allows repeated splits on the same variable within individual trees. As a result, while both notions satisfy symmetry properties, $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$ more closely matches the structural flexibility inherent in regression trees and is therefore the more appropriate notion of variation in this context.

Learnability Beyond $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$

We have argued that XGBoost is expected to effectively estimate functions in the class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$. In particular, if $f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ and the complexity measure $V^{d,s}_{\infty-\mathrm{xgb}}(f^*)$ can be treated as a constant, then the idealized XGBoost estimator—defined as a solution to the XGBoost optimization problem—achieves the curse-of-dimensionality-avoiding rate $n^{-2/3}$, up to logarithmic factors.

A natural follow-up question is what happens when f^* lies outside $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$. A simple example of such a function is

$$f^*(\mathbf{x}) := \mathbf{1}(x_1 + \cdots + x_d \ge 0, \mathbf{x} \in [-1,1]^d). \tag{5.23}$$

It can be shown that this function does not belong to $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$. One way to see this is that f^* has infinite Hardy–Krause variation; see, e.g., [113, Proposition 17].

For functions f^* lying outside $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$, in light of Theorems 17 and 18 of Section 5.3, it is natural to conjecture that the risk of the idealized XGBoost estimator takes the form

$$\inf_V \Bigg((V+1)^{\beta} \cdot n^{-2/3} (\log n)^{\gamma} + \inf_{\substack{f_0 \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}} \\ V^{d,s}_{\infty-\mathrm{xgb}}(f_0) \le V}} \|f_0 - f^*\|^2_{p_0,2} \Bigg), \tag{5.24}$$

for some constants $\beta \in [2/3, 2]$ and $\gamma \in [2(\bar{s}-1)/3, 4(\bar{s}-1)/3]$, where $\bar{s} = \min(s, d)$. The upper bounds on β and γ follow from the risk upper bound in Theorem 17, while the lower bounds are expected from the minimax lower bound in Theorem 18. For simplicity, we suppress the dependence on other parameters, such as d, s, and the distributions of the covariates and error terms.

If we assume $\beta = 2/3$ and ignore the logarithmic factor $(\log n)^\gamma$, then (5.24) reduces to

$$\inf_V \left((V+1)^{2/3} \cdot n^{-2/3} + \inf_{\substack{f_0 \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}} \\ V^{d,s}_{\infty-\mathrm{xgb}}(f_0) \le V}} \|f_0 - f^*\|^2_{p_0,2} \right).$$

For additional simplicity, suppose that p_0 is the uniform density on $[-1,1]^d$. Then, for the function (5.23), one can show that for sufficiently large V,

$$\inf_{\substack{f_0 \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}} \\ V^{d,s}_{\infty-\mathrm{xgb}}(f_0) \le V}} \|f_0 - f^*\|^2_{p_0,2} = \Omega(V^{-1/(d-1)}).$$

Consequently, even in this most favorable scenario (with $\beta = 2/3$), the convergence rate of the idealized XGBoost estimator for this f^* is no faster than

$$\inf_V \left((V+1)^{2/3} \cdot n^{-2/3} + V^{-1/(d-1)} \right) \asymp n^{-2/(2d+1)}.$$

Unlike the curse-of-dimensionality-avoiding rate achieved when $f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ with bounded $V^{d,s}_{\infty-\mathrm{xgb}}(f^*)$, the above rate deteriorates rapidly as the dimension d increases. This suggests that while XGBoost may still achieve consistency under misspecification, it is not well suited for estimating functions that lie far outside the class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$.

L^1 vs L^2 Regularization

We have focused on the L^1 penalty for XGBoost, as defined in (5.1). As mentioned in the chapter introduction, XGBoost implementations also commonly employ a squared L^2 penalty, in which $\|\mathbf{w}_k\|_1$ is replaced by $\|\mathbf{w}_k\|_2^2$. More generally, for any $p \ge 1$, we may define

$$V^{d,s}_{\mathrm{xgb}}(f; p) := \inf \left\{ \sum_k \|\mathbf{w}_k\|_p^p \right\}, \quad \text{for } f \in \mathcal{F}^{d,s}_{\mathrm{st}},$$

where $\|\cdot\|_p$ denotes the usual L^p norm, and as in (5.1), the infimum is taken over all representations of f as a finite sum of right-continuous regression trees of depth at most s. However, this variation functional yields a meaningful regularization penalty only when $p = 1$. Specifically, when $p > 1$, the penalty becomes degenerate, as shown by the following result. This degeneracy explains why we restrict attention to the L^1 penalty in this chapter.

Lemma 7. *Suppose $p > 1$. Then, $V^{d,s}_{xgb}(f; p) = 0$ for every $f \in \mathcal{F}^{d,s}_{st}$.*

In practice, XGBoost operates on the function class $\mathcal{F}_{\text{st}}^{d,s}(K)$ consisting of functions of the form $\sum_{k=1}^{K} f_k$, where each f_k is a regression tree with right-continuous splits and depth $\leq s$. Here, K is a fixed finite number that is typically selected via cross-validation. The distinction between $\mathcal{F}_{\text{st}}^{d,s}$ and $\mathcal{F}_{\text{st}}^{d,s}(K)$ is that the former allows an arbitrary number of trees, whereas the latter restricts attention to ensembles of at most K trees.

Within this restricted class $\mathcal{F}_{\text{st}}^{d,s}(K)$, we can define a truncated version of the penalty by

$$V_{\text{xgb}}^{d,s}(f; p, K) := \inf \left\{ \sum_{k=1}^{K} \|\mathbf{w}_k\|_p^p \right\},$$

where the infimum is now taken over all representations of $f \in \mathcal{F}_{\text{st}}^{d,s}(K)$ as a sum of at most K regression trees of depth at most s. This modified penalty is likely well defined for all $p \geq 1$, including $p = 2$. However, it is theoretically cumbersome due to its rigid dependence on the hyperparameter K. Specifically, this formulation does not admit a meaningful limit as $K \to \infty$. Moreover, because K is data-dependent in practice, it is unnatural to treat it as a fixed number.

For these reasons, we focus exclusively on the case $p = 1$, as this choice provides a stable regularization penalty that generalizes naturally to continuum tree ensembles and avoids the vanishing-penalty issues inherent to norms with $p > 1$ in the absence of a fixed tree count.

Analysis of the Iterative Algorithm Used by XGBoost

Our analysis focuses on the statistical behavior of solutions to the regularized optimization problem (5.2) that XGBoost is designed to approximate. We do not study whether the greedy tree-boosting algorithm employed in practice achieves the same rates of convergence over the class $\mathcal{F}_{\infty-\text{st}}^{d,s}$, and establishing such guarantees remains an important open problem. Some recent progress has been made in the analysis of greedy tree-building algorithms; see, for example, [139, 32, 101, 83, 141].

Despite this limitation, our results remain directly relevant to the practice of XGBoost. By characterizing the behavior of the target optimization problem, our theory provides a principled benchmark for what XGBoost can achieve under favorable optimization. In particular, the results clarify when dimension-free rates are attainable and when intrinsic approximation barriers arise due to misspecification. This perspective helps disentangle statistical limitations—stemming from the expressiveness of the tree ensemble and its associated regularization—from algorithmic limitations of the greedy boosting procedure itself, thereby offering a coherent framework for interpreting the empirical successes and failures of XGBoost in practice.

Chapter 6

Conclusion

In this thesis, we studied multivariate adaptive regression splines (MARS) and extreme gradient boosting (XGBoost) through the infinite-dimensional lasso framework (1.5). This framework is universal in the sense that it can, in principle, be applied to any other regression method whose estimator is constructed as a finite linear combination of its own atoms as in (1.2). As illustrated by the cases of MARS and XGBoost, it can serve as a useful tool for analyzing machine learning methods whose theoretical properties remain poorly understood, often because their estimators are produced by greedy construction procedures. In some cases, as with MARS, it may lead to a variant of the original method that is statistically more principled and theoretically more tractable, while still providing indirect insight into the original method. In other cases, as with XGBoost, it may bring a better understanding of the underlying function classes and optimization problems, thereby clarifying the types of functions to which the method is well suited.

However, it should also be noted that, although the framework itself is universal, most of the results presented in this thesis are specific to MARS and XGBoost. Characterizations in terms of smoothness, as well as connections to mixed partial derivatives, need not arise for other methods. Likewise, the computational and theoretical results obtained for MARS and XGBoost, such as reductions to finite-dimensional optimization problems and nearly dimension-free rates of convergence, may not readily transfer to other settings. Most of the proof techniques developed for these results rely heavily on the particular forms of their atoms. For these reasons, studying other methods through the infinite-dimensional lasso framework would likely lead to an entirely different collection of results, possibly requiring substantially different ideas and proof techniques.

Moreover, studying machine learning methods through the infinite-dimensional lasso framework has fundamental limitations. The framework does not take into account the algorithms actually used by each method. This simplification is precisely what makes rigorous analysis possible and enables us to gain some perspective on methods for which little was previously understood theoretically. However, greedy algorithms are often one of the key drivers of the empirical success of many machine learning methods. Therefore, studying such methods without taking their algorithms into account can provide only a partial understanding in many cases. To fully understand these methods, it is essential to study them together with proper consideration of their algorithmic aspects. We thus believe that, for MARS and XGBoost as well as many other machine learning methods, studying their algorithms is an important task—though often much more challenging—for achieving a more complete understanding of these methods.

In this thesis, we also studied totally concave regression, which can be obtained as an instance of the sign-constrained variant of (1.5) with the MARS atoms. More generally, for any machine learning method that is compatible with the infinite-dimensional lasso framework (1.5), one may consider a corresponding shape-constrained regression method obtained in the same manner. However, unlike the case of totally concave regression, the resulting method need not always correspond to a meaningful or interpretable shape constraint on the underlying function. In many cases, it may lack an independent motivation in terms of controlling shapes of practical interest, of the kind present in totally concave regression. It would therefore be interesting to investigate whether other widely used regression methods give rise, through this construction, to shape-constrained regression methods associated with natural and useful classes of functions.

Bibliography

[1] Christoph Aistleitner and Josef Dick. "Functions of bounded variation, signed measures, and a general Koksma–Hlawka inequality". In: *Acta Arithmetica* 167.2 (2015), pp. 143–171.

[2] Dennis Amelunxen et al. "Living on the edge: Phase transitions in convex programs with random data". In: *Information and Inference: A Journal of the IMA* 3.3 (2014), pp. 224–294.

[3] Shiri Artstein et al. "On convexified packing and entropy duality". In: *Geometric and Functional Analysis* 14.5 (2004), pp. 1134–1141.

[4] Francis Bach. "Breaking the curse of dimensionality with convex neural networks". In: *Journal of Machine Learning Research* 18.19 (2017), pp. 1–53.

[5] Gábor Balázs. "Convex Regression: Theory, Practice, and Applications". PhD thesis. University of Alberta, 2016.

[6] Gábor Balázs, András György, and Csaba Szepesvári. "Near-optimal max-affine estimators for convex regression". In: *Artificial Intelligence and Statistics*. PMLR. 2015, pp. 56–64.

[7] Pierre C Bellec. "Sharp oracle inequalities for least squares estimators in shape restricted regression". In: *Annals of Statistics* 46.2 (2018), pp. 745–780.

[8] David Benkeser and Mark van der Laan. "The highly adaptive lasso estimator". In: *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016, pp. 689–696.

[9] Maïtine Bergounioux. "Mathematical analysis of a inf-convolution model for image processing". In: *Journal of Optimization Theory and Applications* 168.1 (2016), pp. 1–21.

[10] Dimitris Bertsimas and Nishanth Mundru. "Sparse convex regression". In: *INFORMS Journal on Computing* 33.1 (2021), pp. 262–279.

[11] Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. "Atomic norm denoising with applications to line spectral estimation". In: *IEEE Transactions on Signal Processing* 61.23 (2013), pp. 5987–5999.

[12] Herman J Bierens and Donna K Ginther. "Integrated conditional moment testing of quantile regression models". In: *Empirical Economics* 26.1 (2001), pp. 307–324.

[13] Ron Blei, Fuchang Gao, and Wenbo Li. "Metric entropy of high dimensional distributions". In: *Proceedings of the American Mathematical Society* 135.12 (2007), pp. 4009–4018.

[14] Vadim Borisov et al. "Deep neural networks and tabular data: A survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 35.6 (2022), pp. 7499–7519.

[15] Claire Boyer et al. "On representer theorems and convex regularization". In: *SIAM Journal on Optimization* 29.2 (2019), pp. 1260–1281.

[16] Kristian Bredies and Marcello Carioni. "Sparsity of solutions for variational inverse problems with finite-dimensional data". In: *Calculus of Variations and Partial Differential Equations* 59.1 (2020), p. 14.

[17] Kristian Bredies and Martin Holler. "Higher-order total variation approaches and generalisations". In: *Inverse Problems* 36.12 (2020), p. 123001.

[18] Kristian Bredies and Hanna Katriina Pikkarainen. "Inverse problems in spaces of measures". In: *ESAIM: Control, Optimisation and Calculus of Variations* 19.1 (2013), pp. 190–218.

[19] Leo Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[20] Leo Breiman. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3 (2001), pp. 199–231.

[21] Leo Breiman and Jerome H Friedman. "Estimating optimal transformations for multiple regression and correlation". In: *Journal of the American Statistical Association* 80.391 (1985), pp. 580–598.

[22] Thomas F Brooks, D Stuart Pope, and Michael A Marcolini. *Airfoil self-noise and prediction.* Tech. rep. L-16528. NASA, 1989.

[23] Hans-Joachim Bungartz and Michael Griebel. "Sparse grids". In: *Acta Numerica* 13 (2004), pp. 147–269.

[24] Emmanuel J Candès and Carlos Fernandez-Granda. "Towards a mathematical theory of super-resolution". In: *Communications on Pure and Applied Mathematics* 67.6 (2014), pp. 906–956.

[25] Antonin Chambolle and Pierre-Louis Lions. "Image recovery via total variation minimization and related problems". In: *Numerische Mathematik* 76.2 (1997), pp. 167–188.

[26] Sabyasachi Chatterjee. "An improved global risk bound in concave regression". In: *Electronic Journal of Statistics* 10.1 (2016), pp. 1608–1629.

[27] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. "On risk bounds in isotonic and other shape restricted regression problems". In: *Annals of Statistics* 43.4 (2015), pp. 1774–1800.

[28] Sourav Chatterjee. "A new perspective on least squares under convex constraint". In: *Annals of Statistics* 42.6 (2014), pp. 2340–2381.

[29] Tianqi Chen and Carlos Guestrin. "XGBoost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794.

[30] Xia Chen. "Quadratic functionals and small ball probabilities for the m-fold integrated Brownian motion". In: *Annals of Probability* 31.2 (2003), pp. 1052–1077.

[31] Yining Chen and Richard J Samworth. "Generalized additive and index models with shape constraints". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78.4 (2016), pp. 729–754.

[32] Chien-Ming Chi et al. "Asymptotic properties of high-dimensional random forests". In: *Annals of Statistics* 50.6 (2022), pp. 3415–3438.

[33] Laurent Condat. "Atomic norm minimization for decomposition into complex exponentials and optimal transport in Fourier domain". In: *Journal of Approximation Theory* 258 (2020), p. 105456.

[34] Paulo Cortez et al. "Modeling wine preferences by data mining from physicochemical properties". In: *Decision Support Systems* 47.4 (2009), pp. 547–553.

[35] Yohann De Castro and Fabrice Gamboa. "Exact reconstruction using Beurling minimal extrapolation". In: *Journal of Mathematical Analysis and Applications* 395.1 (2012), pp. 336–354.

[36] Yohann De Castro et al. "Exact solutions to super resolution on semi-algebraic domains in higher dimensions". In: *IEEE Transactions on Information Theory* 63.1 (2016), pp. 621–630.

[37] Quentin Denoyelle, Vincent Duval, and Gabriel Peyré. "Support recovery for sparse super-resolution of positive measures". In: *Journal of Fourier Analysis and Applications* 23.5 (2017), pp. 1153–1194.

[38] Quentin Denoyelle et al. "The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy". In: *Inverse Problems* 36.1 (2020), p. 014001.

[39] XGBoost Developers. *XGBoost 3.0.5 documentation*. `https://xgboost.readthedocs.io/en/stable/parameter.html`. Accessed 2025-09-14. 2025.

[40] David L Donoho et al. "High-dimensional data analysis: The curses and blessings of dimensionality". In: *AMS Math Challenges Lecture* 1.2000 (2000), p. 32.

[41] David L Donoho and Jain M Johnstone. "Ideal spatial adaptation by wavelet shrinkage". In: *Biometrika* 81.3 (1994), pp. 425–455.

[42] Lutz Dumbgen, Sandra Freitag, and Geurt Jongbloed. "Consistency of concave regression with an application to current-status data". In: *Mathematical Methods of Statistics* 13.1 (2004), pp. 69–81.

[43] Dinh Dũng, Vladimir Temlyakov, and Tino Ullrich. *Hyperbolic Cross Approximation.* Advanced Courses in Mathematics - CRM Barcelona. Cham: Birkhäuser, 2018.

[44] Thomas Dunker et al. "Metric entropy of integration operators and small ball probabilities for the Brownian sheet". In: *Journal of Approximation Theory* 101.1 (1999), pp. 63–77.

[45] Vincent Duval and Gabriel Peyré. "Exact support recovery for sparse spikes deconvolution". In: *Foundations of Computational Mathematics* 15.5 (2015), pp. 1315–1355.

[46] Billy Fang, Adityanand Guntuboyina, and Bodhisattva Sen. "Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation". In: *Annals of Statistics* 49.2 (2021), pp. 769–792.

[47] Axel Flinth and Pierre Weiss. "Exact solutions of infinite dimensional total-variation regularized problems". In: *Information and Inference: A Journal of the IMA* 8.3 (2019), pp. 407–443.

[48] Michael S Floater. "A chain rule for multivariate divided differences". In: *BIT Numerical Mathematics* 50 (2010), pp. 577–586.

[49] Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications.* Second. John Wiley & Sons, 1999.

[50] Yoav Freund and Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139.

[51] Jerome H Friedman. *Fast MARS.* Tech. rep. LCS110. Department of Statistics, Stanford University, 1993.

[52] Jerome H Friedman. "Greedy function approximation: A gradient boosting machine". In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232.

[53] Jerome H Friedman. "Multivariate adaptive regression splines". In: *Annals of Statistics* 19.1 (1991), pp. 1–67.

[54] Sorin G Gal. *Shape-Preserving Approximation by Real and Complex Polynomials.* Boston, MA: Birkhäuser, 2010.

[55] Fuchang Gao. "Bracketing entropy of high dimensional distributions". In: *High Dimensional Probability VI.* Vol. 66. Progress in Probability. Basel: Birkhäuser, 2013, pp. 3–17.

[56] Fuchang Gao. "Entropy estimate for k-monotone functions via small ball probability of integrated Brownian motions". In: *Electronic Communications in Probability* 13 (2008), pp. 121–130.

[57] Fuchang Gao and Jon A Wellner. "Entropy of convex functions on $\mathbb{R}^d$". In: *Constructive Approximation* 46 (2017), pp. 565–592.

[58] Sara van de Geer. *Applications of Empirical Process Theory*. Vol. 6. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2000.

[59] Sara van de Geer. *Estimation and Testing Under Sparsity: École d'Été de Probabilités de Saint-Flour XLV-2015*. Vol. 2159. Lecture Notes in Mathematics. Cham: Springer, 2016.

[60] Giordano Giambartolomei. "The Karhunen–Loève theorem". PhD thesis. Università di Bologna, 2016.

[61] Evarist Giné, Rafał Latała, and Joel Zinn. "Exponential and moment inequalities for U-statistics". In: *High Dimensional Probability II*. Vol. 47. Progress in Probability. Boston, MA: Birkhäuser, 2000, pp. 13–38.

[62] Loukas Grafakos. *Classical Fourier Analysis*. Third. Graduate Texts in Mathematics. New York, NY: Springer, 2014.

[63] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?" In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 507–520.

[64] Piet Groeneboom and Geurt Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Vol. 38. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2014.

[65] Adityanand Guntuboyina and Bodhisattva Sen. "Covering numbers for convex functions". In: *IEEE Transactions on Information Theory* 59.4 (2012), pp. 1957–1965.

[66] Adityanand Guntuboyina and Bodhisattva Sen. "Global risk bounds and adaptation in univariate convex regression". In: *Probability Theory and Related Fields* 163.1 (2015), pp. 379–411.

[67] Adityanand Guntuboyina and Bodhisattva Sen. "Nonparametric shape-restricted regression". In: *Statistical Science* 33.4 (2018), pp. 568–594.

[68] Adityanand Guntuboyina et al. "Adaptive risk bounds in univariate total variation denoising and trend filtering". In: *Annals of Statistics* 48.1 (2020), pp. 205–229.

[69] Qiyang Han and Jon A Wellner. "Convergence rates of least squares regression estimators with heavy-tailed errors". In: *Annals of Statistics* 47.4 (2019), pp. 2286–2319.

[70] Qiyang Han and Jon A Wellner. "Multivariate convex regression: Global risk bounds and adaptation". In: *arXiv preprint arXiv:1601.06844* (2016).

[71] Qiyang Han and Jon A Wellner. "Robustness of shape-restricted regression estimators: An envelope perspective". In: *arXiv preprint arXiv:1805.02542* (2018).

[72] D. L. Hanson and Gordon Pledger. "Consistency in concave regression". In: *Annals of Statistics* 4.6 (1976), pp. 1038–1050.

[73] Clifford Hildreth. "Point estimates of ordinates of concave functions". In: *Journal of the American Statistical Association* 49.267 (1954), pp. 598–619.

[74] Eugene Isaacson and Herbert Bishop Keller. *Analysis of Numerical Methods*. Revised. Dover, 1994.

[75] Jiro Iwanaga et al. "Estimating product-choice probabilities from recency and frequency of page views". In: *Knowledge-Based Systems* 99 (2016), pp. 157–167.

[76] Iain M Johnstone. "Gaussian estimation: Sequence and wavelet models". In: *Manuscript, September 2019* (2019). available at `https://imjohnstone.su.domains/GE_09_16_19.pdf`.

[77] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in Neural Information Processing Systems* 30 (2017).

[78] Dohyeong Ki, Billy Fang, and Adityanand Guntuboyina. "MARS via LASSO". In: *arXiv preprint arXiv:2111.11694v1* (2021). URL: `https://arxiv.org/pdf/2111.11694v1`.

[79] Dohyeong Ki, Billy Fang, and Adityanand Guntuboyina. "MARS via LASSO". In: *Annals of Statistics* 52.3 (2024), pp. 1102–1126.

[80] Dohyeong Ki and Adityanand Guntuboyina. "Totally concave regression". In: *Journal of the American Statistical Association* (2026), pp. 1–12.

[81] Dohyeong Ki and Adityanand Guntuboyina. "What functions does XGBoost learn?" In: *arXiv preprint arXiv:2601.05444* (2026).

[82] Seung-Jean Kim et al. "ℓ_1 trend filtering". In: *SIAM review* 51.2 (2009), pp. 339–360.

[83] Jason M Klusowski and Peter M Tian. "Large scale prediction with decision trees". In: *Journal of the American Statistical Association* 119.545 (2024), pp. 525–537.

[84] Arun K Kuchibhotla and Rohit K Patra. "On least squares estimation under heteroscedastic and heavy-tailed errors". In: *Annals of Statistics* 50.1 (2022), pp. 277–302.

[85] Arun K Kuchibhotla, Rohit K Patra, and Bodhisattva Sen. "Semiparametric efficiency in convexity constrained single-index model". In: *Journal of the American Statistical Association* 118.541 (2023), pp. 272–286.

[86] Vladimir N Kulikov and Hendrik P Lopuhaä. "The behavior of the NPMLE of a decreasing density near the boundaries of the support". In: *Annals of Statistics* 34.2 (2006), pp. 742–768.

[87] Timo Kuosmanen. "Representation theorem for convex nonparametric least squares". In: *Econometrics Journal* 11.2 (2008), pp. 308–325.

[88] Gil Kur, Yuval Dagan, and Alexander Rakhlin. "Optimality of maximum likelihood for log-concave density estimation and bounded convex regression". In: *arXiv preprint arXiv:1903.05315* (2019).

[89] Gil Kur et al. "Convex regression in multidimensions: Suboptimality of least squares estimators". In: *Annals of Statistics* 52.6 (2024), pp. 2791–2815.

[90] Mark J van der Laan, David Benkeser, and Weixin Cai. "Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso". In: *International Journal of Biostatistics* 19.1 (2023), pp. 261–289.

[91] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Vol. 23. A Series of Modern Surveys in Mathematics. Berlin, Heidelberg: Springer, 1991.

[92] Thomas Lemieux. "The 'Mincer equation' thirty years after schooling, experience, and earnings". In: *Jacob Mincer a Pioneer of Modern Labor Economics*. Boston, MA: Springer, 2006, pp. 127–145.

[93] Aleksandr Sergeevich Leonov. "On the total variation for functions of several variables and a multidimensional analog of Helly's selection principle". In: *Mathematical Notes* 63.1 (1996), pp. 61–71.

[94] Wenbo V Li and Werner Linde. "Approximation, metric entropy and small ball estimates for Gaussian measures". In: *Annals of Probability* 27.3 (1999), pp. 1556–1578.

[95] Zhiqiang Liao et al. "Boundary problem and overfitting reduction in convex regression". In: *European Journal of Operational Research* (2026).

[96] Mikhail Anatolevich Lifshits. *Gaussian Random Functions*. Vol. 322. Mathematics and Its Applications. Springer, Dordrecht, 1995.

[97] Eunji Lim. "On convergence rates of convex regression in multiple dimensions". In: *INFORMS Journal on Computing* 26.3 (2014), pp. 616–628.

[98] Eunji Lim and Peter W Glynn. "Consistency of multidimensional convex regression". In: *Operations Research* 60.1 (2012), pp. 196–208.

[99] Stanisław Łojasiewicz. *An Introduction to the Theory of Real Functions*. Chichester: John Wiley & Sons, 1988.

[100] Enno Mammen and Sara van de Geer. "Locally adaptive regression splines". In: *Annals of Statistics* 25.1 (1997), pp. 387–413.

[101] Rahul Mazumder and Haoyue Wang. "On the convergence of CART under sufficient impurity decrease condition". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 57754–57782.

[102] Rahul Mazumder et al. "A computational framework for multivariate convex regression and its variants". In: *Journal of the American Statistical Association* 114.525 (2019), pp. 318–331.

[103] Nicolai Meinshausen. "Sign-constrained least squares estimation for high-dimensional regression". In: *Electronic Journal of Statistics* 7 (2013), pp. 1607–1631.

[104] David Meyer, Friedrich Leisch, and Kurt Hornik. "The support vector machine under test". In: *Neurocomputing* 55.1-2 (2003), pp. 169–186.

[105] Mary C Meyer. "A framework for estimation and inference in generalized additive models with shape and order restrictions". In: *Statistical Science* 33.4 (2018), pp. 595–614.

[106] Mary C Meyer. "Semi-parametric additive constrained regression". In: *Journal of Nonparametric Statistics* 25.3 (2013), pp. 715–730.

[107] Jacob Mincer. *Schooling, Experience and Earnings*. New York, NY: Columbia University Press, 1974.

[108] Kevin M Murphy and Finis Welch. "The structure of wages". In: *Quarterly Journal of Economics* 107.1 (1992), pp. 285–326.

[109] Arkadi Nemirovski. "Topics in Nonparametric Statistics". In: *Lectures on Probability Theory and Statistics: École d'Été de Probabilités de Saint-Flour XXVIII-1998*. Vol. 1738. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer, 2000.

[110] Constantin Niculescu and Lars-Erik Persson. *Convex Functions and Their Applications*. First. CMS Books in Mathematics. New York, NY: Springer, 2006.

[111] Greg Ongie et al. "A function space view of bounded norm infinite width ReLU nets: The multivariate case". In: *arXiv preprint arXiv:1910.01635* (2019).

[112] Francesco Ortelli and Sara van de Geer. "Tensor denoising with trend filtering". In: *Mathematical Statistics and Learning* 4 (2022), pp. 87–142.

[113] Art B Owen. "Multidimensional variation for quasi-Monte Carlo". In: *Contemporary Multivariate Analysis and Design of Experiments*. Vol. 2. Series in Biostatistics. Hackensack, NJ: World Scientific, 2005, pp. 49–74.

[114] Rahul Parhi and Robert D Nowak. "Banach space representer theorems for neural networks and ridge splines". In: *Journal of Machine Learning Research* 22.43 (2021), pp. 1–40.

[115] Rahul Parhi and Robert D Nowak. "Near-minimax optimal estimation with shallow ReLU neural networks". In: *IEEE Transactions on Information Theory* 69.2 (2022), pp. 1125–1140.

[116] Rahul Parhi and Robert D Nowak. "What kinds of functions do deep neural networks learn? Insights from variational spline theory". In: *SIAM Journal on Mathematics of Data Science* 4.2 (2022), pp. 464–489.

[117] Mert Pilanci and Tolga Ergen. "Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7695–7705.

[118] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. "The geometry of off-the-grid compressed sensing". In: *Foundations of Computational Mathematics* 23.1 (2023), pp. 241–327.

[119] Clarice Poon and Gabriel Peyré. "Multidimensional sparse super-resolution". In: *SIAM Journal on Mathematical Analysis* 51.1 (2019), pp. 1–44.

[120] Tiberiu Popoviciu. "Sur quelques propriétés des fonctions d'une ou de deux variables réelles". PhD thesis. Institutul de Arte Grafice" Ardealul, 1933.

[121] Daniel Potts and Michael Schmischke. "Interpretable approximation of high-dimensional data". In: *SIAM Journal on Mathematics of Data Science* 3.4 (2021), pp. 1301–1323.

[122] Daniel Potts and Michael Schmischke. "Learning multivariate functions with low-dimensional structures using polynomial bases". In: *Journal of Computational and Applied Mathematics* 403 (2022), p. 113821.

[123] Liudmila Prokhorenkova et al. "CatBoost: Unbiased boosting with categorical features". In: *Advances in Neural Information Processing Systems* 31 (2018).

[124] Natalya Pya and Simon N Wood. "Shape constrained additive models". In: *Statistics and Computing* 25 (2015), pp. 543–559.

[125] Saharon Rosset et al. "ℓ_1 regularization in infinite dimensional feature spaces". In: *International Conference on Computational Learning Theory*. Vol. 4539. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2007, pp. 544–558.

[126] Walter Rudin. *Real and Complex Analysis*. Third. New York: McGraw–Hill, 1987.

[127] Veeranjaneyulu Sadhanala et al. "Multivariate trend filtering for lattice data". In: *Annals of Statistics* 52.5 (2024), pp. 2400–2430.

[128] Richard J Samworth. "Recent progress in log-concave density estimation". In: *Statistical Science* 33.4 (2018), pp. 493–509.

[129] Richard J Samworth and Bodhisattva Sen. "Editorial: Special issue on 'nonparametric inference under shape constraints'". In: *Statistical Science* 33.4 (2018), pp. 469–472.

[130] Alejandro Schuler, Yi Li, and Mark van der Laan. "Lassoed tree boosting". In: *arXiv preprint arXiv:2205.10697* (2022).

[131] Emilio Seijo and Bodhisattva Sen. "Nonparametric least squares estimation of a multivariate convex regression function". In: *Annals of Statistics* 39.3 (2011), pp. 1633–1657.

[132] Simon Setzer and Gabriele Steidl. "Variational methods with higher order derivatives in image processing". In: *Approximation* 12 (2008), pp. 360–386.

[133] Simon Setzer, Gabriele Steidl, and Tanja Teuber. "Infimal convolution regularizations with discrete ℓ_1-type functionals". In: *Communications in Mathematical Sciences* 9.3 (2011), pp. 797–827.

[134] Ravid Shwartz-Ziv and Amitai Armon. "Tabular data: Deep learning is not all you need". In: *Information Fusion* 81 (2022), pp. 84–90.

[135] Jonathan W Siegel and Jinchao Xu. “Characterization of the variation spaces corresponding to shallow neural networks”. In: *Constructive Approximation* 57.3 (2023), pp. 1109–1132.

[136] Martin Slawski and Matthias Hein. “Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization”. In: *Electronic Journal of Statistics* 7 (2013), p. 3004.

[137] Gabriele Steidl, Stephan Didas, and Julia Neumann. “Splines in higher order TV regularization”. In: *International Journal of Computer Vision* 70.3 (2006), pp. 241–255.

[138] Jiayang Sun and Michael Woodroofe. “Adaptive smoothing for a penalized NPMLE of a non-increasing density”. In: *Journal of Statistical Planning and Inference* 52.2 (1996), pp. 143–159.

[139] Vasilis Syrgkanis and Manolis Zampetakis. “Estimation and inference with trees and forests in high dimensions”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 3453–3454.

[140] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems*. Second. Vol. 60. A Series of Modern Surveys in Mathematics. Cham: Springer, 2021.

[141] Yan Shuo Tan, Jason M Klusowski, and Krishnakumar Balasubramanian. “Statistical-computational trade-offs for recursive adaptive partitioning estimators”. In: *arXiv preprint arXiv:2411.04394* (2024).

[142] Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. “Near minimax line spectral estimation”. In: *IEEE Transactions on Information Theory* 61.1 (2014), pp. 499–512.

[143] Vladimir Temlyakov. *Multivariate Approximation*. Vol. 32. Cambridge Monographs on Applied and Computational Mathematics. Cambridge: Cambridge University Press, 2018.

[144] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[145] Ryan J Tibshirani. “Adaptive piecewise polynomial estimation via trend filtering”. In: *Annals of Statistics* 42.1 (2014), pp. 285–323.

[146] Shawn W Ulrick. “Measuring the returns to education nonparametrically”. In: *Applied Economics Letters* 14.13 (2007), pp. 1005–1011.

[147] Michael Unser. “Ridges, neural networks, and the Radon transform”. In: *Journal of Machine Learning Research* 24.37 (2023), pp. 1–33.

[148] Michael Unser, Julien Fageot, and John Paul Ward. "Splines are universal solutions of linear inverse problems with generalized TV regularization". In: *SIAM Review* 59.4 (2017), pp. 769–793.

[149] Aad W van der Vaart. *Asymptotic Statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2000.

[150] Aad W van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. New York, NY: Springer, 1996.

[151] Aad W van der Vaart and J Harry van Zanten. "Reproducing kernel Hilbert spaces of Gaussian priors". In: *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*. Vol. 3. Institute of Mathematical Statistics Collections. Institute of Mathematical Statistics, 2008, pp. 200–222.

[152] Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Vol. 48. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2019.

[153] Michael Woodroofe and Jiayang Sun. "A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing". In: *Statistica Sinica* 3.2 (1993), pp. 501–515.

[154] I-Cheng Yeh. "Modeling of strength of high-performance concrete using artificial neural networks". In: *Cement and Concrete Research* 28.12 (1998), pp. 1797–1808.

Appendix A

MARS

A.1 Results of Section 3.1 for the case $d = s = 2$

According to the definition (3.3), $\mathcal{F}^{2,2}_{\infty-\text{mars}}$ consists of all functions of the form

$$\begin{aligned} f(x_1, x_2) = c &+ \int_{[0,1)} (x_1 - t_1)_+ \, d\nu_1(t_1) + \int_{[0,1)} (x_2 - t_2)_+ \, d\nu_2(t_2) \\ &+ \int_{[0,1)^2} (x_1 - t_1)_+ (x_2 - t_2)_+ \, d\nu_{12}(t_1, t_2) \end{aligned}$$

for $c \in \mathbb{R}$ and finite signed measures ν_1, ν_2, and ν_{12} (on $[0, 1), [0, 1)$, and $[0, 1)^2$, respectively). On the other hand, according to the characterization in Proposition 1, $\mathcal{F}^{2,2}_{\infty-\text{mars}}$ consists of all functions of the form

$$f(x_1, x_2) = c + \int_0^{x_1} g_1(t_1) \, dt_1 + \int_0^{x_2} g_2(t_2) \, dt_2 + \int_{[0,x_1]\times[0,x_2]} g_{12}(t_1, t_2) \, d(t_1, t_2)$$

for $c \in \mathbb{R}$ and real-valued functions g_1, g_2, and g_{12} (on $[0, 1], [0, 1]$, and $[0, 1]^2$, respectively) which have finite Hardy–Krause variation and satisfy the one-sided continuity conditions stated in Proposition 1. The functions g_1, g_2, and g_{12} can be considered as particular derivatives of f, and they are denoted $D^1 f$, $D^2 f$, and $D^{12} f$. Hence, we can write the estimator $\hat{f}^{2,2}_{n,V}$ alternatively as

$$\hat{f}^{2,2}_{n,V} \in \operatorname*{argmin}_f \left\{ \sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)}) \big)^2 : \mathrm{HK}_{\mathbf{0}}(D^{12} f) + \mathrm{TV}(D^1 f) + \mathrm{TV}(D^2 f) \le V \right\}.$$

Moreover, according to Proposition 2, the functions g_1, g_2, and g_{12} coincide with another type of derivatives $D_1 f$, $D_2 f$, and $D_{12} f$ of f almost everywhere (with respect to the Lebesgue measure). For smooth functions f, the formula (3.13) for $d = s = 2$ becomes

$$V_{\mathrm{mars}}(f) = \int_0^1 \int_0^1 \left|f^{(2,2)}(x_1, x_2)\right| dx_1 dx_2 + \int_0^1 \left|f^{(2,1)}(x_1, 0)\right| dx_1 + \int_0^1 \left|f^{(2,0)}(x_1, 0)\right| dx_1 + \int_0^1 \left|f^{(1,2)}(0, x_2)\right| dx_2 + \int_0^1 \left|f^{(0,2)}(0, x_2)\right| dx_2.$$

This can be viewed as a second-order version of the Hardy–Krause variation:

$$\mathrm{HK}_{\mathbf{0}}(f) = \int_0^1 \int_0^1 \left|f^{(1,1)}(x_1, x_2)\right| dx_1 dx_2 + \int_0^1 \left|f^{(1,0)}(x_1, 0)\right| dx_1 + \int_0^1 \left|f^{(0,1)}(0, x_2)\right| dx_2.$$

A.2 Brief Descriptions of Datasets

- *Earnings Dataset.* This dataset is from `ex1029` in the R library Sleuth3 (see also [12]). It includes weekly earnings of 25,437 full-time male workers in 1987, along with their years of education and years of experience. The original dataset also has a few more indicator variables, but we do not include them in our analysis. We predict the logarithm of weekly earnings from years of education and years of experience.

- *Airfoil Self-Noise Dataset.* This dataset is available at `https://archive.ics.uci.edu/dataset/291/airfoil+self+noise` (see also [22]). It contains the results of aerodynamic and acoustic tests on (NACA 0012) airfoils of different sizes, conducted in an anechoic wind tunnel. The tests measured the scaled sound pressure level of the self-noise of the airfoils at various wind tunnel speeds and angles of attack. This dataset has 1503 observations and six variables including the scaled sound pressure level. Here, the scaled sound pressure level is predicted from the other five variables describing test settings.

- *Abalone Dataset.* This dataset is from abalone in the R library AppliedPredictiveModeling, and it is also available at `https://archive.ics.uci.edu/dataset/1/abalone`. This dataset is composed of the physical measurements of 4177 abalones. We aim to predict the number of rings each abalone has from seven physical measurements: length, diameter, height, whole weight, and the weight of meat, gut, and shell. The number of rings is of interest because it can be used for estimating age. Here, we exclude from our analysis the indicator variable for the sex of each abalone.

- *Concrete Dataset.* This dataset is originally from [154] and currently available at `https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength`. It contains the compressive strength of 1030 concrete samples together with their ages and the quantity of seven components, such as cement, fly ash, and water. In our analysis, we have the compressive strength as a response variable and the other eight variables as covariates.

- *Ozone Dataset.* This dataset is from ozone in the R library faraway (see also [21]). It consists of 330 observations of atmospheric ozone concentration level and nine other meteorological variables, including temperature, wind speed, humidity, and visibility, measured in the Los Angeles Basin in 1976. Here, we predict atmospheric ozone concentration level from the other meteorological variables.

- *Wine Dataset.* This dataset was first introduced in [34], and it is currently available at `https://archive.ics.uci.edu/dataset/186/wine+quality`. It comprises the quality of wines (*vinho verde*) from the Minho region of Portugal and eleven other attributes of wines exhibiting their physicochemical properties; for example, alcohol content, density, pH, and the quantity of citric acid. The quality of wines was graded on a 10-point scale, with 0 indicating very bad and 10 indicating excellent. This dataset has two parts: one for red wines and the other for white wines. There are 1599 red wine data and 4898 white wine data. In our analysis, we fit models separately for red and white wines, predicting the quality of wines from the physicochemical variables.

Appendix B

Totally Concave Regression

B.1 Axially Concave Regression

Let $\hat{f}_n^d$ denote the least squares estimator over the class of axially concave functions on $[0,1]^d$:

$$\hat{f}_n^d \in \underset{f}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)})\big)^2 : f \text{ is axially concave} \right\}. \tag{B.1}$$

In this section, we provide details for the computation of $\hat{f}_n^d$ that we deployed in Section 4.4.

For each $j \in [d]$, write

$$\big\{0, x_j^{(1)}, \dots, x_j^{(n)}, 1\big\} = \big\{v_0^{(j)}, \dots, v_{n_j}^{(j)}\big\}$$

where $0 = v_0^{(j)} < \cdots < v_{n_j}^{(j)} = 1$. Then, for each $i = 1, \dots, n$, there exists $\mathbf{b}(i) = (b_1(i), \dots, b_d(i)) \in I_0 := \prod_{j=1}^d \{0, \dots, n_j\}$ such that $x_j^{(i)} = v_{b_j(i)}^{(j)}$ for every $j \in [d]$. Also, for each real-valued function f on $[0,1]^d$, let $\boldsymbol{\theta}_f \in \mathbb{R}^{|I_0|}$ denote the vector of evaluations of f at $(v_{l_1}^{(1)}, \dots, v_{l_d}^{(d)})$ for $(l_1, \dots, l_d) \in I_0$; that is,

$$(\theta_f)_{l_1,\dots,l_d} = f\big(v_{l_1}^{(1)}, \dots, v_{l_d}^{(d)}\big)$$

for $(l_1, \dots, l_d) \in I_0$. Note that $f(\mathbf{x}^{(i)}) = (\theta_f)_{\mathbf{b}(i)}$ for every $i = 1, \dots, n$.

Suppose f is an axially concave function on $[0,1]^d$. By the the axial concavity of f,

$$\begin{aligned} &\frac{1}{v_{l_j+1}^{(j)} - v_{l_j}^{(j)}} \cdot \big((\theta_f)_{l_1,\dots,l_{j-1},l_j+1,l_{j+1},\dots,l_d} - (\theta_f)_{l_1,\dots,l_{j-1},l_j,l_{j+1},\dots,l_d}\big) \\ &\qquad \geq \frac{1}{v_{l_j+2}^{(j)} - v_{l_j+1}^{(j)}} \cdot \big((\theta_f)_{l_1,\dots,l_{j-1},l_j+2,l_{j+1},\dots,l_d} - (\theta_f)_{l_1,\dots,l_{j-1},l_j+1,l_{j+1},\dots,l_d}\big) \end{aligned} \tag{B.2}$$

for all $0 \leq l_k \leq n_k$ $(k \neq j)$ and $0 \leq l_j \leq n_j - 2$. Conversely, if $\boldsymbol{\theta} \in \mathbb{R}^{|I_0|}$ satisfies (B.2), then there exists an axially concave function f on $[0,1]^d$ such that $\boldsymbol{\theta}_f = \boldsymbol{\theta}$. This can be readily proved as follows.

Let f denote the function on $[0,1]^d$ defined by

$$f(x_1,\dots,x_d) = \frac{1}{\prod_{j=1}^d (v^{(j)}_{l_j+1} - v^{(j)}_{l_j})} \sum_{\boldsymbol{\delta} \in \{0,1\}^d} \prod_{j=1}^d \big\{(1-\delta_j)\big(v^{(j)}_{l_j+1} - x_j\big) + \delta_j\big(x_j - v^{(j)}_{l_j}\big)\big\} \cdot \theta_{l_1+\delta_1,\dots,l_d+\delta_d} \tag{B.3}$$

for $(x_1,\dots,x_d) \in \prod_{j=1}^d [v^{(j)}_{l_j}, v^{(j)}_{l_j+1}]$ for each $(l_1,\dots,l_d) \in \prod_{j=1}^d \{0,1,\dots,n_j-1\}$. Observe that $\boldsymbol{\theta}_f = \boldsymbol{\theta}$. Also, observe that, for every $j \in [d]$ and for every fixed $(x_1,\dots,x_{j-1},x_{j+1},\dots,x_d) \in [0,1]^{d-1}$, the function $g_j(t) := f(x_1,\dots,x_{j-1},t,x_{j+1},\dots,x_d)$ is continuous and piecewise affine in t over $[0,1]$. In fact, the function f is a continuous rectangular piecewise multi-affine function (introduced in Section 4.3), interpolating the points $((v^{(1)}_{l_1},\dots,v^{(d)}_{l_d}), \theta_{l_1,\dots,l_d})$ for $(l_1,\dots,l_d) \in I_0$.

If $(x_1,\dots,x_j,x_{j+1},\dots,x_d) \in \prod_{k\neq j}[v^{(k)}_{l_k}, v^{(k)}_{l_k+1}]$, then the slope of g_j at t is given by

$$\begin{aligned}
&\frac{1}{\prod_{k\neq j}(v^{(k)}_{l_k+1} - v^{(k)}_{l_k})} \cdot \sum_{\boldsymbol{\delta}\in\{0,1\}^{d-1}} \prod_{k\neq j} \big\{(1-\delta_k)\big(v^{(k)}_{l_k+1} - x_k\big) + \delta_k\big(x_k - v^{(k)}_{l_k}\big)\big\} \\
&\quad \cdot \frac{1}{v^{(j)}_{l_j+1} - v^{(j)}_{l_j}} \cdot \big(\theta_{l_1+\delta_1,\dots,l_{j-1}+\delta_{j-1},l_j+1,l_{j+1}+\delta_{j+1},\dots,l_d+\delta_d} - \theta_{l_1+\delta_1,\dots,l_{j-1}+\delta_{j-1},l_j,l_{j+1}+\delta_{j+1},\dots,l_d+\delta_d}\big)
\end{aligned}$$

for $t \in (v^{(j)}_{l_j}, v^{(j)}_{l_j+1})$. By (B.2), this slope is decreasing in t, implying that g_j is concave. Hence, f is axially concave, as desired.

Now, consider the following finite-dimensional least squares problem:

$$\hat{\boldsymbol{\theta}}^d_n \in \operatorname*{argmin}_{\boldsymbol{\theta}} \big\{\|\mathbf{y} - \mathbf{M}\boldsymbol{\theta}\|_2 : \boldsymbol{\theta} \in \mathbb{R}^{|I_0|} \text{ satisfies (B.2)}\big\}, \tag{B.4}$$

where $\mathbf{y} = (y_1,\dots,y_n)$, and $\mathbf{M}$ is the $n \times |I_0|$ matrix with

$$M_{i,(l_1,\dots,l_d)} = \begin{cases} 1 & \text{if } (l_1,\dots,l_d) = \mathbf{b}(i) \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1,\dots,n$ and for $(l_1,\dots,l_d) \in I_0$. Recall that $\boldsymbol{\theta}_f$ satisfies (B.2) for every axially concave function f, and there always exists some axially concave function f on $[0,1]^d$ such that $\boldsymbol{\theta}_f = \boldsymbol{\theta}$ for every $\boldsymbol{\theta} \in \mathbb{R}^{|I_0|}$ satisfying (B.2). Also, recall that $f(\mathbf{x}^{(i)}) = (\boldsymbol{\theta}_f)_{\mathbf{b}(i)}$ for every $i = 1,\dots,n$. It thus follows that if we define f as in (B.3) using $\hat{\boldsymbol{\theta}}^d_n$ instead of $\boldsymbol{\theta}$, then such f solves the original least squares problem (B.1). Hence, once we solve (B.4), we can construct the least squares estimator for axially concave regression using (B.3). This is also what we used in Section 4.4.

However, it should be noted that solving (B.4) can be computationally challenging when the dimension d is not small. The number of components of the vector $\boldsymbol{\theta}$ in (B.4) is $\prod_{j=1}^d (n_j+1)$, which can grow as large as $(n+2)^d$ in the worst case. Recall that, for each $j \in [d]$, $n_j + 1$ is the number of unique values observed in the j^{th} component of $\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}$. This computational burden is also the reason why axially concave regression was not considered for the housing price dataset and the 401(k) dataset in Section 4.4. Therefore, developing more efficient methods for computing $\hat{f}_n^d$ or computationally feasible variants is crucial for making axially concave regression practical. Exploring these alternatives would be an interesting direction for future research.

B.2 Additional Plots

In this section, we provide additional plots for the earnings dataset. Figure B.1 shows the proportion of repetitions in which each model is among the two best-performing models (Plot A) and the worst-performing model (Plot B). This figure is similar to Figure 4.2 (the only differences are best vs. top-two and worst vs. bottom-two). Like the plots in Figure 4.2, those in Figure B.1 indicate that for small sample sizes, Mincer's and Murphy and Welch's parametric models, as well as additive concave regression, are often better options. For large sample sizes, totally and axially concave regressions generally perform best. More interestingly, for moderate sample sizes, totally concave regression consistently ranks among the top two models, whereas axially concave regression frequently becomes the worst-performing model. These observations reaffirm the findings from Figure 4.2.

Figure B.2 displays the empirical cumulative distributions of model ranks across 100 splits into training and test sets, for four different training data proportions. As in the cumulative rank plots presented in Section 4.4, ranks are determined by mean squared errors on the test sets, with rank 1 indicating the best-performing model and rank 5 the worst. Plot A corresponds to a training data proportion of $0.9/2^7$, representing a small sample size regime, while Plot D corresponds to $0.9/2^1$, representing a large sample size regime. Plots B and C correspond to intermediate training data proportions of $0.9/2^5$ and $0.9/2^3$, respectively.

An ideal method is one whose rank distribution function is everywhere larger than the others. Plot A shows that Mincer's model is the ideal method in the small sample size regime. As the sample size increases to the level of Plot B, Murphy and Welch's model becomes the most ideal. In both cases, axially concave regression is the worst choice. When the sample size reaches that of Plot C, totally concave regression is clearly the best option. Finally, for the largest sample size in Plot D, totally and axially concave regressions are the two best options that are comparable.

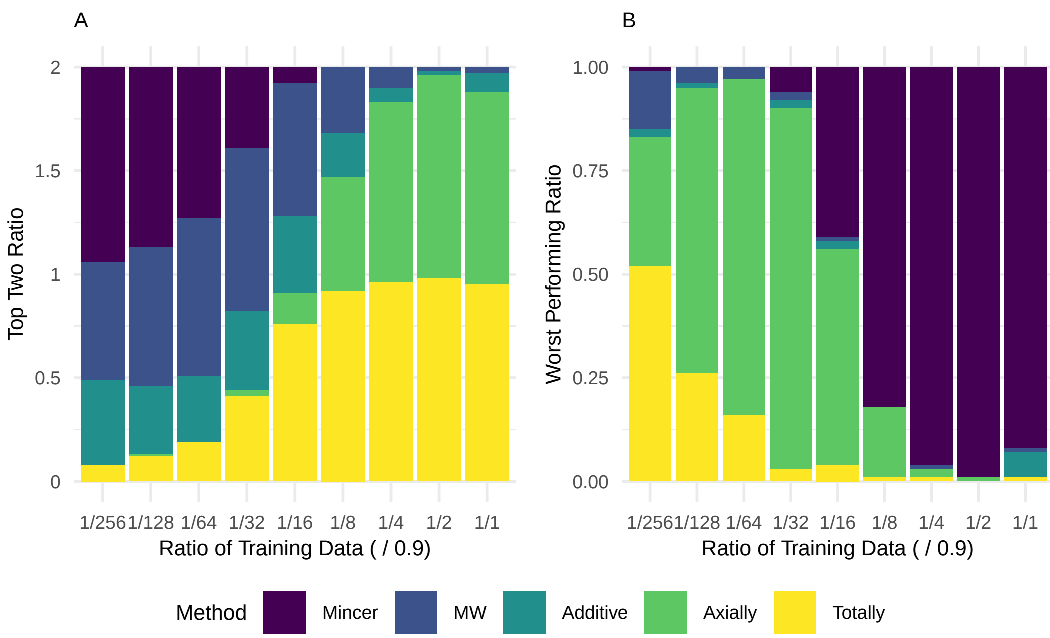

Figure B.1: Top two and worst performing ratios of each method for various sub-sample sizes for the earnings dataset. The x-axis and y-axis represent the proportion of training data and the proportion of repetitions (out of 100) where (A) each model is one of the two best-performing models and (B) each model is the worst-performing model.

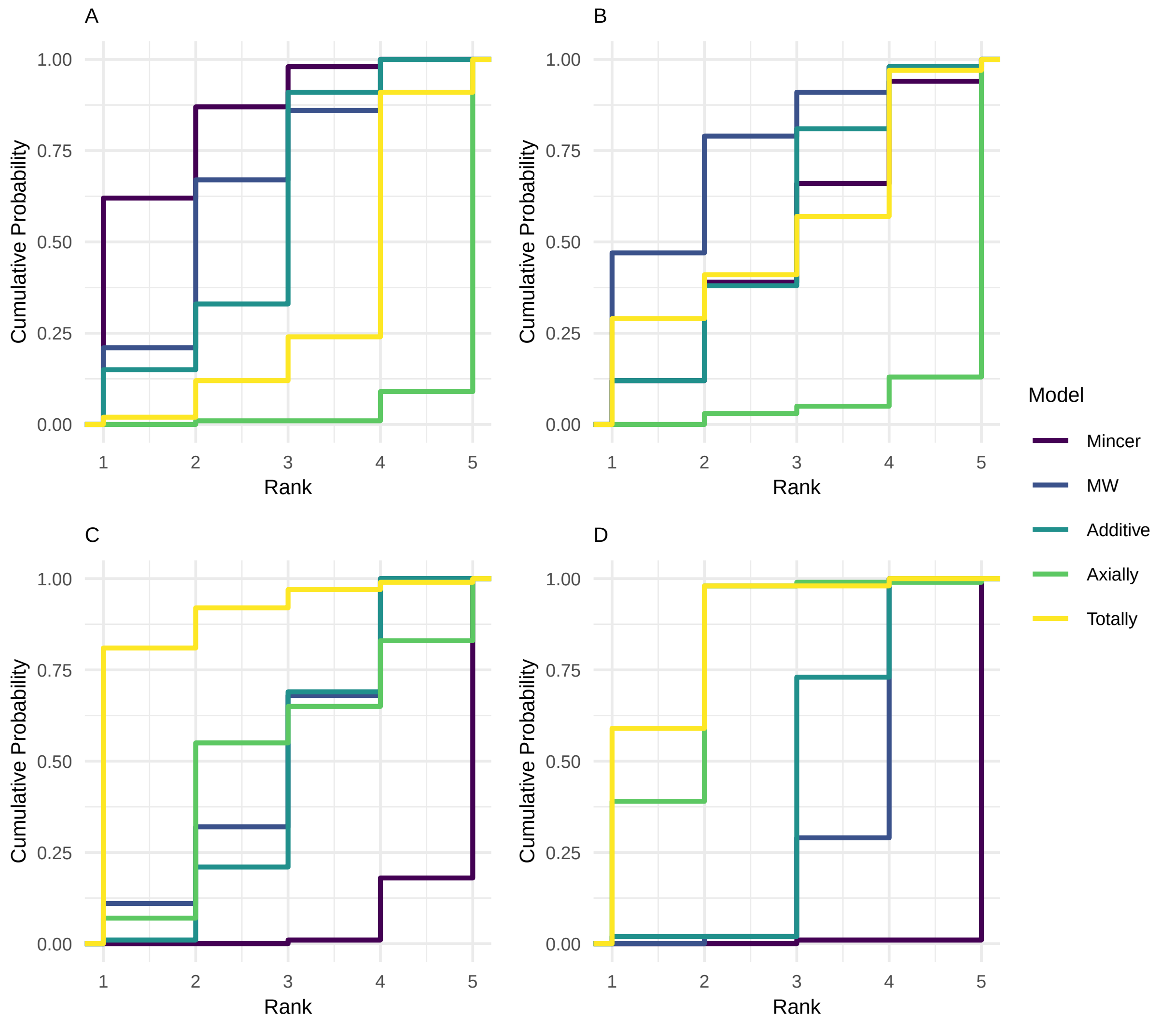

Figure B.2: Empirical cumulative distributions of model ranks across 100 splits into training and test sets, for various proportions of training data for the earnings dataset. The training data proportions are (A) $0.9/2^7$, (B) $0.9/2^5$, (C) $0.9/2^3$, and (D) $0.9/2^1$. Ranks are determined by mean squared errors on the test sets.

Appendix C

Proofs for MARS

In the proofs presented in this and the following chapters, C (and sometimes c) denotes a universal constant. When a constant depends on specific variables, this dependence will be indicated through subscripts. The value of a constant may vary from one line to another, even if the same notation is used. Also, for simplicity, the exact values of constants are often left unspecified.

C.1 Uniqueness of Representation (3.3)

Here we prove the following lemma, which shows that the constant c and the signed measures ν_S appearing in the representation (3.3) of $f \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$ are uniquely determined by f.

Lemma 8. *Suppose c and b are real numbers and $\{\nu_S\}$ and $\{\mu_S\}$ are collections of finite signed measures as in* (3.3)*. If $f_{c,\{\nu_S\}} \equiv f_{b,\{\mu_S\}}$ on $[0,1]^d$, then $c = b$ and $\nu_S = \mu_S$ for all $S \subseteq [d]$ with $0 < |S| \le s$.*

Proof of Lemma 8. By the assumption,

$$c + \sum_{0<|S|\le s} \int_{[0,1)^{|S|}} \prod_{j\in S}(x_j - t_j)_+ \, d\nu_S(t_j, j \in S) = b + \sum_{0<|S|\le s} \int_{[0,1)^{|S|}} \prod_{j\in S}(x_j - t_j)_+ \, d\mu_S(t_j, j \in S)$$

for all $(x_1, \dots, x_d) \in [0,1]^d$. Substituting $x_1 = \cdots = x_d = 0$, we obtain $c = b$. Suppose for some positive integer $m \le s$, we have $\nu_S = \mu_S$ for all $S \subseteq [d]$ with $0 < |S| \le m-1$. Substituting $x_{m+1} = \cdots = x_d = 0$ and eliminating from both sides the components that are already known to be equal, we obtain

$$\int_{[0,1)^m} \prod_{j=1}^{m}(x_j - t_j)_+ \, d\nu_{[m]}(\mathbf{t}) = \int_{[0,1)^m} \prod_{j=1}^{m}(x_j - t_j)_+ \, d\mu_{[m]}(\mathbf{t}), \tag{C.1}$$

where $[m] = \{1, \ldots, m\}$. Note that

$$\int_{[0,1)^m} \prod_{j=1}^{m} (x_j - t_j)_+ \, d\nu_{[m]}(\mathbf{t}) = \int_{\prod_{j=1}^m [0,x_j)} \prod_{j=1}^{m} (x_j - t_j) \, d\nu_{[m]}(\mathbf{t})$$

$$= \int_{\prod_{j=1}^m [0,x_j)} \int_{\prod_{j=1}^m (t_j,x_j)} 1 \, d\mathbf{s} \, d\nu_{[m]}(\mathbf{t}) = \int_{\prod_{j=1}^m (0,x_j)} \int_{\prod_{j=1}^m [0,s_j)} 1 \, d\nu_{[m]}(\mathbf{t}) \, d\mathbf{s}.$$

Thus, equation (C.1) can be written as

$$\int_{\prod_{j=1}^m (0,x_j)} \int_{\prod_{j=1}^m [0,s_j)} d\nu_{[m]}(\mathbf{t}) \, d\mathbf{s} = \int_{\prod_{j=1}^m (0,x_j)} \int_{\prod_{j=1}^m [0,s_j)} d\mu_{[m]}(\mathbf{t}) \, d\mathbf{s}.$$

By the Lebesgue differentiation theorem (see, e.g., [126, Theorem 7.10]), it follows that

$$\int_{\prod_{j=1}^m [0,x_j)} d\nu_{[m]}(\mathbf{t}) = \int_{\prod_{j=1}^m [0,x_j)} d\mu_{[m]}(\mathbf{t}) \tag{C.2}$$

almost everywhere with respect to the Lebesgue measure.

For a subset R of $[0,1)^m$, we call it an axis-aligned rectangle if it is of the form

$$R = \prod_{j=1}^{m} I_j,$$

where for each j, $I_j = (v_j, w_j), [v_j, w_j], [v_j, w_j)$, or $(v_j, w_j]$ for some real numbers v_j and w_j. Now, let $\mathcal{P}$ be the collection of all axis-aligned rectangles in $[0,1)^m$ and $\mathcal{Q}$ be the collection of all Borel sets $E \subseteq [0,1)^m$ for which

$$\nu_{[m]}(E) = \mu_{[m]}(E)$$

Using equation (C.2), we can show that $\mathcal{P} \subseteq \mathcal{Q}$. Also, it is clear that $\mathcal{P}$ is a π-system, and $\mathcal{Q}$ is a λ-system (or Dynkin system) on $[0,1)^m$. Therefore, by Dynkin's π-λ theorem, we have $\sigma(\mathcal{P}) \subseteq \mathcal{Q}$. Since $\sigma(\mathcal{P})$ is indeed the collection of all Borel sets in $[0,1)^m$, we can conclude that $\nu_{[m]} = \mu_{[m]}$. Consequently, induction argument completes the proof. □

C.2 Proof of Proposition 1

In our proof of Proposition 1, we use the following variant of Theorem 1 in Chapter 2. The proof of this lemma is given in Appendix C.14.

Lemma 9. *Suppose a real-valued function g on $[0,1]^m$ has finite $HK_{\mathbf{0}}$ variation and is right-continuous on $[0,1]^m$. Assume also that g is left-continuous at each point $(x_1, \ldots, x_m) \in$*

$[0,1]^m \setminus [0,1)^m$ *with respect to all the* j^{th} *coordinates where* $x_j = 1$. *Then, there exists a unique finite signed (Borel) measure* ν *on* $[0,1)^m$ *such that*

$$g(x_1, \dots, x_m) = \nu\Big(\prod_{j \in [m]} [0, x_j] \cap [0,1)^m \Big) \quad \text{for every } (x_1, \dots, x_m) \in [0,1]^m. \tag{C.3}$$

Also, if a real-valued function g *defined on* $[0,1]^m$ *is of the form* (C.3) *for some finite signed measure* ν *on* $[0,1)^m$, *then* g *has finite* $HK_{\mathbf{0}}$ *variation and*

$$|\nu|([0,1)^m \setminus \{\mathbf{0}\}) = HK_{\mathbf{0}}(g).$$

Proof of Proposition 1. Suppose $f \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$. Then, there exist $c \in \mathbb{R}$ and a collection of finite signed measures $\{\nu_S : S \subseteq [d], 0 < |S| \le s\}$ as in (3.3) such that $f = f_{c,\{\nu_S\}}$. For each $S \subseteq [d]$ with $0 < |S| \le s$, let g_S be the function on $[0,1]^{|S|}$ defined by

$$g_S(x_j, j \in S) = \nu_S\Big(\prod_{j \in S} [0, x_j] \cap [0,1)^{|S|} \Big)$$

for $(x_j, j \in S) \in [0,1]^{|S|}$. Clearly, g_S is right-continuous and left-continuous at each point $(x_j, j \in S) \in [0,1]^{|S|} \setminus [0,1)^{|S|}$ with respect to all the j^{th} coordinates where $x_j = 1$. Also, by Lemma 9, we can see that g_S has finite $\text{HK}_{\mathbf{0}}$ variation. Moreover, we have

$$\begin{aligned}
f(x_1, \dots, x_d) &= c + \sum_{0<|S|\le s} \int_{[0,1)^{|S|}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S) \\
&= c + \sum_{0<|S|\le s} \int_{\prod_{j \in S}[0,x_j)} \prod_{j \in S} (x_j - t_j) \, d\nu_S(t_j, j \in S) \\
&= c + \sum_{0<|S|\le s} \int_{\prod_{j \in S}[0,x_j)} \int_{\prod_{j \in S}[t_j,x_j)} 1 \, d(s_j, j \in S) \, d\nu_S(t_j, j \in S) \\
&= c + \sum_{0<|S|\le s} \int_{\prod_{j \in S}[0,x_j)} \int_{\prod_{j \in S}[0,s_j]} 1 \, d\nu_S(t_j, j \in S) \, d(s_j, j \in S) \\
&= c + \sum_{0<|S|\le s} \int_{\prod_{j \in S}[0,x_j)} \nu_S\Big(\prod_{j \in S} [0, s_j] \Big) \, d(s_j, j \in S) \\
&= c + \sum_{0<|S|\le s} \int_{\prod_{j \in S}[0,x_j]} g_S(s_j, j \in S) \, d(s_j, j \in S)
\end{aligned}$$

for all $(x_1, \dots, x_d) \in [0,1]^d$. Thus, f can be represented as in (3.12) with $c \in \mathbb{R}$ and the collection of functions $\{g_S : S \subseteq [d], 0 < |S| \le s\}$ that has the desired properties.

Conversely, suppose a real-valued function f on $[0,1]^d$ is of the form

$$f(x_1, \dots, x_d) = c + \sum_{0<|S|\le s} \int_{\prod_{j \in S}[0,x_j]} g_S(t_j, j \in S) \, d(t_j, j \in S)$$

for some $c \in \mathbb{R}$ and for some collection of functions $\{g_S : S \subseteq [d], 0 < |S| \leq s\}$, where for each $S \subseteq [d]$ with $0 < |S| \leq s$, $g_S : [0,1]^{|S|} \to \mathbb{R}$ has finite $\mathrm{HK}_{\mathbf{0}}$ variation, is right-continuous, and is left-continuous at each point $(x_j, j \in S) \in [0,1]^{|S|} \setminus [0,1)^{|S|}$ with respect to all the j^{th} coordinates where $x_j = 1$. For each $S \subseteq [d]$ with $0 < |S| \leq s$, by Lemma 9 again, there exists a finite signed measure ν_S on $[0,1)^{|S|}$ such that

$$g_S(x_j, j \in S) = \nu_S\Big(\prod_{j \in S}[0, x_j] \cap [0,1)^{|S|}\Big) \quad \text{for } (x_j, j \in S) \in [0,1]^{|S|},$$

and

$$|\nu_S|([0,1)^{|S|} \setminus \{\mathbf{0}\}) = \mathrm{HK}_{\mathbf{0}}(g_S).$$

Hence, the function f can be written as

$$\begin{aligned} f(x_1, \dots, x_d) &= c + \sum_{0<|S|\leq s} \int_{\prod_{j \in S}[0,x_j]} g_S(t_j, j \in S)\, d(t_j, j \in S) \\ &= c + \sum_{0<|S|\leq s} \int_{\prod_{j \in S}[0,x_j)} \nu_S\Big(\prod_{j \in S}[0, t_j]\Big)\, d(t_j, j \in S) \\ &= c + \sum_{0<|S|\leq s} \int_{[0,1)^{|S|}} \prod_{j \in S}(x_j - s_j)_+ \, d\nu_S(s_j, j \in S), \end{aligned}$$

which means that $f \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$. Moreover, we can represent the complexity measure of f as

$$V_{\text{mars}}(f) = \sum_{0<|S|\leq s} |\nu_S|([0,1)^{|S|} \setminus \{\mathbf{0}\}) = \sum_{0<|S|\leq s} \mathrm{HK}_{\mathbf{0}}(g_S).$$

□

C.3 Proof of Proposition 2

Here, we prove a stronger statement that directly implies Proposition 2. We show that for f in condition (3.12), the derivatives of the interval functions associated with f exist and are equal to g_S in the same condition almost everywhere (with respect to the Lebesgue measure). We start by introducing interval functions, their derivatives, and some related concepts. We refer the reader to [99] for a more detailed introduction to them.

Two closed real intervals are said to be adjoining if they have a common endpoint and do not share interior points. For example, $[0,1]$ and $[1,2]$ are two adjoining closed intervals. Similarly, we say that two m-dimensional axis-aligned (closed) rectangles

$$\prod_{j=1}^{m}[x_j, y_j] \quad \text{and} \quad \prod_{j=1}^{m}[z_j, w_j]$$

are adjoining if $[x_j, y_j]$ and $[z_j, w_j]$ are adjoining for some $j_0 \in [m]$ and $[x_j, y_j] = [z_j, w_j]$ for all $j \neq j_0$. For instance, $[0,1] \times [0,1]$ and $[0,1] \times [1,2]$ are two adjoining two-dimensional axis-aligned rectangles. Note that for two adjoining axis-aligned rectangles, their union is also an axis-aligned rectangle. A real-valued function F defined on a class of m-dimensional axis-aligned (closed) rectangles is then called an additive interval function if

$$F(R \cup R') = F(R) + F(R')$$

for every pair of m-dimensional adjoining axis-aligned rectangles R and R'.

For an axis-aligned rectangle $\prod_{j=1}^m [x_j, y_j]$,

$$\left| \prod_{j=1}^m [x_j, y_j] \right| := (y_1 - x_1) \cdots (y_m - x_m)$$

denotes its volume. Also, the rectangle is referred to as a cube if it is the product of closed intervals of equal length, i.e., $y_1 - x_1 = \cdots = y_m - x_m$. The upper and lower derivatives of an additive interval function F at $\mathbf{x} \in \mathbb{R}^m$ are then defined by

$$\overline{D}F(\mathbf{x}) = \limsup_{\substack{C:\text{cube, } C \ni \mathbf{x} \\ |C| \to 0}} \frac{F(C)}{|C|} \quad \text{and} \quad \underline{D}F(\mathbf{x}) = \liminf_{\substack{C:\text{cube, } C \ni \mathbf{x} \\ |C| \to 0}} \frac{F(C)}{|C|}.$$

If the two derivatives coincide, we say that the derivative of F at $\mathbf{x}$ exists and define it as the common value

$$DF(\mathbf{x}) = \lim_{\substack{C:\text{cube, } C \ni \mathbf{x} \\ |C| \to 0}} \frac{F(C)}{|C|}.$$

The Lebesgue differentiation theorem (see, e.g., [126, Theorem 7.10]) can be stated in terms of interval functions and their derivatives as follows. The following theorem plays a key role in identifying g_S in condition (3.12).

Theorem 19 (Theorem 7.1.7 of [99]). *Let f be an integrable real-valued function defined on an axis-aligned rectangle R_0. Suppose that F is the additive interval function defined by*

$$F(R) = \int_R f(\mathbf{x}) d\mathbf{x}$$

for axis-aligned rectangles $R \subseteq R_0$. Then, it follows that $DF = f$ almost everywhere (with respect to the Lebesgue measure) on R_0.

We now define, for a real-valued function f on $[0,1]^d$, the interval functions associated with f. For each nonempty subset $S \subseteq [d]$, denote by $\Delta_S f$ the additive interval function which is defined on the class of axis-aligned (closed) rectangles contained in $[0,1]^{|S|}$ and which is defined by

$$\Delta_S f\left(\prod_{j \in S} [x_j, y_j] \right) = \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S} \delta_j} f\big(\widetilde{\mathbf{x}\mathbf{y}}_{\boldsymbol{\delta}} \big),$$

where $\widetilde{\mathbf{xy}}_{\boldsymbol{\delta}}$ is the d-dimensional vector with

$$\left(\widetilde{\mathbf{xy}}_{\boldsymbol{\delta}}\right)_j = \begin{cases} \delta_j x_j + (1-\delta_j) y_j & \text{if } j \in S \\ 0 & \text{otherwise} \end{cases}$$

for $j \in [d]$. For example, if $d = 3$ and $S = \{1, 3\}$, $\Delta_{13} f$ is defined by

$$\Delta_{13} f([x_1, y_1] \times [x_3, y_3]) = f(y_1, 0, y_3) - f(y_1, 0, x_3) - f(x_1, 0, y_3) + f(x_1, 0, x_3).$$

Also, if $S = [d]$, $\Delta_{[d]} f(\prod_{j=1}^d [x_j, y_j])$ coincides with the quasi-volume of f on $\prod_{j=1}^d [x_j, y_j]$ for all $0 \leq x_j < y_j \leq 1$ for $j \in [d]$. Recall the definition of quasi-volume from Section 2.1. We call these $\Delta_S f$ the interval functions associated with f.

We are ready to state and prove the stronger statement describing the relationship between f and g_S in condition (3.12). It is clear from the definitions above that Proposition 2 directly follows from this result.

Proposition 15. *Suppose that condition* (3.12) *holds. Then, for each nonempty subset* $S \subseteq [d]$, $D\Delta_S f = 0$ *if* $|S| > s$, *and* $D\Delta_S f = g_S$ *almost everywhere (with respect to the Lebesgue measure) on* $[0,1]^{|S|}$ *if* $|S| \leq s$.

Proof of Proposition 15. Recall that condition (3.12) assumes that

$$f(x_1, \ldots, x_d) = c + \sum_{0 < |S| \leq s} \int_{\prod_{j \in S} [0, x_j]} g_S(t_j, j \in S) \, d(t_j, j \in S) \quad \text{for } (x_1, \ldots, x_d) \in [0,1]^d,$$

where $c \in \mathbb{R}$ and g_S is a real-valued function on $[0,1]^{|S|}$ for each $S \subseteq [d]$ with $0 < |S| \leq s$. By letting $g_S = 0$ for each $S \subseteq [d]$ with $|S| > s$, we can rewrite it as

$$f(x_1, \ldots, x_d) = c + \sum_{\emptyset \neq S \subseteq [d]} \int_{\prod_{j \in S} [0, x_j]} g_S(t_j, j \in S) \, d(t_j, j \in S)$$

for $(x_1, \ldots, x_d) \in [0,1]^d$. Then, for each nonempty subset $S \subseteq [d]$, we have

$$\begin{aligned}
\Delta_S f\left(\prod_{j \in S} [x_j, y_j]\right) &= \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S} \delta_j} \cdot f\left(\widetilde{\mathbf{xy}}_{\boldsymbol{\delta}}\right) \\
&= \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S} \delta_j} \cdot \sum_{T : \emptyset \neq T \subseteq S} \int_{\prod_{j \in T} [0, \delta_j x_j + (1-\delta_j) y_j]} g_T(t_j, j \in T) \, d(t_j, j \in T) \\
&= \sum_{T : \emptyset \neq T \subseteq S} \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S} \delta_j} \cdot \int_{\prod_{j \in T} [0, \delta_j x_j + (1-\delta_j) y_j]} g_T(t_j, j \in T) \, d(t_j, j \in T) \\
&= \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S} \delta_j} \cdot \int_{\prod_{j \in S} [0, \delta_j x_j + (1-\delta_j) y_j]} g_S(t_j, j \in S) \, d(t_j, j \in S)
\end{aligned}$$

$$= \int_{\prod_{j\in S}[x_j,y_j]} g_S(t_j, j \in S)\, d(t_j, j \in S)$$

for all $0 \leq x_j < y_j \leq 1$ for $j \in S$. Thus, by Theorem 19, the derivative $D\Delta_S f$ exists and equals g_S almost everywhere (with respect to the Lebesgue measure) for each nonempty subset $S \subseteq [d]$. Because we let $g_S = 0$ for $S \subseteq [d]$ with $|S| > s$, this means that $D\Delta_S f = 0$ if $|S| > s$, and $D\Delta_S f = g_S$ almost everywhere if $0 < |S| \leq s$. □

C.4 Proof of Lemma 1

We use the following lemmas in the proof of Lemma 1. Lemma 10 is a consequence of the fundamental theorem of calculus, while Lemma 11 is a version of (2.3) in Chapter 2. We provide proofs of these lemmas right after the proof of Lemma 1.

Lemma 10. *Suppose a real-valued function g defined on $[0,1]^m$ is smooth in the sense that, for every $\mathbf{p} \in \{0,1\}^m$, $g^{(\mathbf{p})}$ exists and is continuous on $[0,1]^m$. For each $\mathbf{p} \in \{0,1\}^m$, let $S_{\mathbf{p}} = \{j \in [m] : p_j = 1\}$ and $\mathbf{t}^{(\mathbf{p})} = (t_j, j \in S_{\mathbf{p}})$. Then, g can be expressed as*

$$g(x_1, \ldots, x_m) = g(0, \ldots, 0) + \sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}} \int_{\prod_{j\in S_{\mathbf{p}}}[0,x_j]} g^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})})\, d\mathbf{t}^{(\mathbf{p})}$$

for $(x_1, \ldots, x_m) \in [0,1]^m$, where $\widetilde{\mathbf{t}}^{(\mathbf{p})}$ is an m-dimensional extension of $\mathbf{t}^{(\mathbf{p})}$ for which

$$\widetilde{t}_j^{(\mathbf{p})} = \begin{cases} t_j & \text{if } j \in S_{\mathbf{p}} \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 11. *Suppose a real-valued function g defined on $[0,1]^m$ is smooth in the sense that, for every $\mathbf{p} \in \{0,1\}^m$, $g^{(\mathbf{p})}$ exists and is continuous on $[0,1]^m$. Then, g has finite $HK_{\mathbf{0}}$ variation and*

$$HK_{\mathbf{0}}(g) = \sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}} \int_{T^{(\mathbf{p})}} |g^{(\mathbf{p})}|. \tag{C.4}$$

Proof of Lemma 1. Suppose f is smooth in the sense that it satisfies the condition in the lemma, and assume that $f^{(\mathbf{p})} = 0$ for every $\mathbf{p} \in \{0,1\}^d$ with $\sum_j p_j > s$. For each $\mathbf{p} \in \{0,1\}^d \setminus \{\mathbf{0}\}$ with $\sum_j p_j \leq s$, let $g_{\mathbf{p}}$ be the function on $[0,1]^{|S_{\mathbf{p}}|}$ defined by

$$g_{\mathbf{p}}(x_j, j \in S_{\mathbf{p}}) = f^{(\mathbf{p})}(\widetilde{\mathbf{x}}^{(\mathbf{p})})$$

for $(x_j, j \in S_{\mathbf{p}}) \in [0,1]^{|S_{\mathbf{p}}|}$. Note that since $f^{(\mathbf{p})}$ is continuous on $T^{(\mathbf{p})}$, $g_{\mathbf{p}}$ is continuous on $[0,1]^{|S_{\mathbf{p}}|}$. Also, since $f^{(\mathbf{q})}$ exists and is continuous on $T^{(\mathbf{p})}$ for every $\mathbf{q} \in J_{S_{\mathbf{p}}}$, we can see that

$g_{\mathbf{p}}^{(\boldsymbol{\delta})}$ exists and is continuous on $[0,1]^{|S_{\mathbf{p}}|}$ for every $\boldsymbol{\delta} \in \{0,1\}^{|S_{\mathbf{p}}|}$. Thus, by Lemma 11, it follows that $g_{\mathbf{p}}$ has finite $\mathrm{HK}_{\mathbf{0}}$ variation and

$$\mathrm{HK}_{\mathbf{0}}(g_{\mathbf{p}}) = \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S_{\mathbf{p}}|} \setminus \{\mathbf{0}\}} \int_{T_{\mathbf{p}}^{(\boldsymbol{\delta})}} |g_{\mathbf{p}}^{(\boldsymbol{\delta})}|,$$

where

$$T_{\mathbf{p}}^{(\boldsymbol{\delta})} := \prod_{j \in S_{\mathbf{p}}} T_j^{(\boldsymbol{\delta})} \quad \text{where} \quad T_j^{(\boldsymbol{\delta})} = \begin{cases} [0,1] & \text{if } \delta_j = \max_k \delta_k \\ \{0\} & \text{otherwise.} \end{cases}$$

Moreover, since $f^{(\mathbf{p})}$ exists and is continuous on $[0,1]^d$ for every $\mathbf{p} \in \{0,1\}^d$, Lemma 10 implies that for every $(x_1, \dots, x_d) \in [0,1]^d$,

$$\begin{aligned} f(x_1, \dots, x_d) &= f(\mathbf{0}) + \sum_{\mathbf{p} \in \{0,1\}^d \setminus \{\mathbf{0}\}} \int_{\prod_{j \in S_{\mathbf{p}}} [0, x_j]} f^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})}) \, d\mathbf{t}^{(\mathbf{p})} \\ &= f(\mathbf{0}) + \sum_{\substack{\mathbf{p} \in \{0,1\}^d \setminus \{\mathbf{0}\} \\ \sum_j p_j \le s}} \int_{\prod_{j \in S_{\mathbf{p}}} [0, x_j]} g_{\mathbf{p}}(t_j, j \in S_{\mathbf{p}}) \, d(t_j, j \in S_{\mathbf{p}}) \end{aligned}$$

For these reasons, by Proposition 1, we can conclude that $f \in \mathcal{F}_{\infty-\mathrm{mars}}^{d,s}$ and that

$$\begin{aligned} V_{\mathrm{mars}}(f) &= \sum_{\substack{\mathbf{p} \in \{0,1\}^d \setminus \{\mathbf{0}\} \\ \sum_j p_j \le s}} \mathrm{HK}_{\mathbf{0}}(g_{\mathbf{p}}) = \sum_{\substack{\mathbf{p} \in \{0,1\}^d \setminus \{\mathbf{0}\} \\ \sum_j p_j \le s}} \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S_{\mathbf{p}}|} \setminus \{\mathbf{0}\}} \int_{T_{\mathbf{p}}^{(\boldsymbol{\delta})}} |g_{\mathbf{p}}^{(\boldsymbol{\delta})}| \\ &= \sum_{\substack{\mathbf{p} \in \{0,1\}^d \setminus \{\mathbf{0}\} \\ \sum_j p_j \le s}} \sum_{\mathbf{q} \in J_{S_{\mathbf{p}}}} \int_{T^{(\mathbf{q})}} |f^{(\mathbf{q})}| = \sum_{0 < |S| \le s} \sum_{\mathbf{q} \in J_S} \int_{T^{(\mathbf{q})}} |f^{(\mathbf{q})}|. \end{aligned}$$

If $s = d$, $V_{\mathrm{mars}}(f)$ can also be simplified as

$$V_{\mathrm{mars}}(f) = \sum_{\emptyset \ne S \subseteq [d]} \sum_{\mathbf{q} \in J_S} \int_{T^{(\mathbf{q})}} |f^{(\mathbf{q})}| = \sum_{\substack{\mathbf{q} \in \{0,1,2\}^d \\ \max_j q_j = 2}} \int_{T^{(\mathbf{q})}} |f^{(\mathbf{q})}|.$$

□

Proof of Lemma 10. Note that for $\mathbf{p} \in \{0,1\}^m \setminus \{\mathbf{0}\}$,

$$\int_{\prod_{j \in S_{\mathbf{p}}} [0, x_j]} g^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})}) \, d\mathbf{t}^{(\mathbf{p})} = \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S_{\mathbf{p}}|}} (-1)^{\sum_{j \in S_{\mathbf{p}}} \delta_j} g\big(\widetilde{\mathbf{x}}_{\boldsymbol{\delta}}^{(\mathbf{p})}\big),$$

where $\widetilde{\mathbf{x}}_{\boldsymbol{\delta}}^{(\mathbf{p})}$ is the m-dimensional vector with

$$\left(\widetilde{\mathbf{x}}_{\boldsymbol{\delta}}^{(\mathbf{p})}\right)_j = \begin{cases} (1-\delta_j)x_j & \text{if } j \in S_{\mathbf{p}} \\ 0 & \text{otherwise} \end{cases}$$

for $j \in [m]$. It can be easily proved by the fundamental theorem of calculus, and we use the smoothness assumption on g here. For example, if $m = 2$ and $\mathbf{p} = (1,1)$,

$$\begin{aligned}
&\int_{[0,x_1]\times[0,x_2]} g^{(1,1)}(t_1,t_2)\, d(t_1,t_2) = \int_0^{x_1}\int_0^{x_2} g^{(1,1)}(t_1,t_2)\, dt_2\, dt_1 \\
&\quad = \int_0^{x_1} \left(g^{(1,0)}(t_1,x_2) - g^{(1,0)}(t_1,0)\right) dt_1 = \int_0^{x_1} g^{(1,0)}(t_1,x_2)\, dt_1 - \int_0^{x_1} g^{(1,0)}(t_1,0)\, dt_1 \\
&\quad = g(x_1,x_2) - g(0,x_2) - g(x_1,0) + g(0,0) = \sum_{\boldsymbol{\delta}\in\{0,1\}^2} (-1)^{\delta_1+\delta_2} \cdot g\left(\widetilde{\mathbf{x}}_{\boldsymbol{\delta}}^{(1,1)}\right),
\end{aligned}$$

and we use that $g, g^{(1,0)}, g^{(0,1)}$, and $g^{(1,1)}$ are continuous on $[0,1]^2$ for this argument. Thus,

$$\begin{aligned}
&g(\mathbf{0}) + \sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}} \int_{\prod_{j\in S_{\mathbf{p}}}[0,x_j]} g^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})})\, d\mathbf{t}^{(\mathbf{p})} \\
&\quad = g(\mathbf{0}) + \sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}} \sum_{\boldsymbol{\delta}\in\{0,1\}^{|S_{\mathbf{p}}|}} (-1)^{\sum_{j\in S_{\mathbf{p}}}\delta_j} \cdot g\left(\widetilde{\mathbf{x}}_{\boldsymbol{\delta}}^{(\mathbf{p})}\right) \\
&\quad = \sum_{\mathbf{p}\in\{0,1\}^m} \sum_{\boldsymbol{\delta}\in\{0,1\}^{|S_{\mathbf{p}}|}} (-1)^{\sum_{j\in S_{\mathbf{p}}}\delta_j} \cdot g\left(\widetilde{\mathbf{x}}_{\boldsymbol{\delta}}^{(\mathbf{p})}\right) = \sum_{\mathbf{p}'\in\{0,1\}^m} g(\widetilde{\mathbf{x}}^{(\mathbf{p}')}) \cdot \sum_{\substack{\mathbf{p}\in\{0,1\}^m \\ \mathbf{p}\geq\mathbf{p}'}} (-1)^{|S_{\mathbf{p}}|-|S_{\mathbf{p}'}|}.
\end{aligned}$$

Here, $\mathbf{p} \geq \mathbf{p}'$ means that $p_j \geq p'_j$ for all $j \in [m]$, and the last equality results from counting the number of $g(\widetilde{\mathbf{x}}^{(\mathbf{p}')})$ for each $\mathbf{p}' \in \{0,1\}^m$. Note that if $\mathbf{p}' \neq \mathbf{1}$,

$$\sum_{\substack{\mathbf{p}\in\{0,1\}^m \\ \mathbf{p}\geq\mathbf{p}'}} (-1)^{|S_{\mathbf{p}}|-|S_{\mathbf{p}'}|} = \sum_{\boldsymbol{\delta}\in\{0,1\}^{|S^c_{\mathbf{p}'}|}} (-1)^{\sum_{j\in S^c_{\mathbf{p}'}}\delta_j} = (1-1)^{|S^c_{\mathbf{p}'}|} = 0,$$

where $S^c_{\mathbf{p}'} = \{j \in [m] : p'_j = 0\}$, and that if $\mathbf{p}' = \mathbf{1}$,

$$\sum_{\substack{\mathbf{p}\in\{0,1\}^m \\ \mathbf{p}\geq\mathbf{p}'}} (-1)^{|S_{\mathbf{p}}|-|S_{\mathbf{p}'}|} = 1.$$

As a result, it follows that

$$g(\mathbf{0}) + \sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}} \int_{\prod_{j\in S_{\mathbf{p}}}[0,x_j]} g^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})})\, d\mathbf{t}^{(\mathbf{p})} = g(\widetilde{\mathbf{x}}^{(\mathbf{1})}) = g(x_1,\ldots,x_m).$$

□

Proof of Lemma 11. First, note that since g is smooth in the sense of Lemma 10, we have

$$g(x_1,\ldots,x_m)=g(\mathbf{0})+\sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}}\int_{\prod_{j\in S_{\mathbf{p}}}[0,x_j]} g^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})})\,d\mathbf{t}^{(\mathbf{p})} \tag{C.5}$$

for all $(x_1,\ldots,x_m)\in[0,1]^m$, where $S_{\mathbf{p}}=\{j\in[m]:p_j=1\}$. For each $\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}$, let $\mu_{\mathbf{p}}$ be the signed measure on $(0,1]^{|S_{\mathbf{p}}|}$ defined by

$$d\mu_{\mathbf{p}}=g^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})})\,d\mathbf{t}^{(\mathbf{p})}.$$

Also, for each $\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}$, consider the projection $\pi_{\mathbf{p}}:\mathbb{R}^m\to\mathbb{R}^{|S_{\mathbf{p}}|}$ for which

$$\pi_{\mathbf{p}}(t_j,j\in[m])=(t_j,j\in S_{\mathbf{p}}),$$

and let $H^{(\mathbf{p})}=\prod_{j=1}^m H_j^{(\mathbf{p})}$ where

$$H_j^{(\mathbf{p})}=\begin{cases}(0,1] & \text{if } p_j=\max_k p_k\\ \{0\} & \text{otherwise.}\end{cases}$$

Next, we patch together these signed measures and form the signed measure μ on $[0,1]^m$ by

$$\mu(E)=g(\mathbf{0})\cdot\mathbf{1}(\mathbf{0}\in E)+\sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}}\mu_{\mathbf{p}}(\pi_{\mathbf{p}}(E\cap H^{(\mathbf{p})}))$$

for Borel sets $E\subseteq[0,1]^m$. Observe that

$$\begin{aligned}|\mu|([0,1]^m)&=|g(\mathbf{0})|+\sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}}|\mu_{\mathbf{p}}|((0,1]^{|S_{\mathbf{p}}|})\\&=|g(\mathbf{0})|+\sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}}\int_{(0,1]^{|S_{\mathbf{p}}|}}|g^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})})|\,d\mathbf{t}^{(\mathbf{p})}=|g(\mathbf{0})|+\sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}}\int_{T^{(\mathbf{p})}}|g^{(\mathbf{p})}|.\end{aligned}\tag{C.6}$$

By (C.5), we also have

$$\begin{aligned}g(x_1,\ldots,x_m)&=g(\mathbf{0})+\sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}}\int_{\prod_{j\in S_{\mathbf{p}}}(0,x_j]} g^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})})\,d\mathbf{t}^{(\mathbf{p})}\\&=g(\mathbf{0})+\sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}}\int_{\prod_{j\in S_{\mathbf{p}}}(0,x_j]}d\mu_{\mathbf{p}}=g(\mathbf{0})+\sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}}\mu_{\mathbf{p}}\left(\prod_{j\in S_{\mathbf{p}}}(0,x_j]\right)\\&=g(\mathbf{0})\cdot\mathbf{1}\left(\mathbf{0}\in\prod_{j=1}^m[0,x_j]\right)+\sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}}\mu_{\mathbf{p}}\left(\pi_{\mathbf{p}}\left(\prod_{j=1}^m[0,x_j]\cap H^{(\mathbf{p})}\right)\right)=\mu\left(\prod_{j=1}^m[0,x_j]\right)\end{aligned}$$

for $(x_1,\ldots,x_m)\in[0,1]^m$. Therefore, Theorem 1 implies that g has finite $\mathrm{HK}_{\mathbf{0}}$ variation and

$$|\mu|([0,1]^m)=\mathrm{HK}_{\mathbf{0}}(g)+|g(\mathbf{0})|.$$

Combining this with equation (C.6), we obtain that

$$\mathrm{HK}_{\mathbf{0}}(g) = \sum_{\mathbf{p}\in\{0,1\}^m\setminus\{\mathbf{0}\}} \int_{T^{(\mathbf{p})}} |g^{(\mathbf{p})}|.$$

□

C.5 Proof of Lemma 2

Although the ideas of our proof of Lemma 2 are simple, the proof is somewhat technical because of the intensive handling of indices. Hence, to deliver the underlying ideas clearly, we first prove the lemma when $d = s = 1$ and then consider general d and s.

Proof of Lemma 2. First, assume that $d = s = 1$. For $l = 0, \ldots, n_1 - 1$, let $\widetilde{\mu}_l$ be the discrete signed measure on $[0,1]$ that is supported on $\{v_l^{(1)}, v_{l+1}^{(1)}\}$ and has masses

$$\widetilde{\mu}_l\big(\{v_l^{(1)}\}\big) = \int_{[v_l^{(1)}, v_{l+1}^{(1)})} \frac{v_{l+1}^{(1)} - t}{v_{l+1}^{(1)} - v_l^{(1)}}\, d\nu_1(t) \quad \text{and} \quad \widetilde{\mu}_l\big(\{v_{l+1}^{(1)}\}\big) = \int_{[v_l^{(1)}, v_{l+1}^{(1)})} \frac{t - v_l^{(1)}}{v_{l+1}^{(1)} - v_l^{(1)}}\, d\nu_1(t).$$

Then, for $x \geq v_{l+1}^{(1)}$,

$$\begin{aligned}
&\int_{[v_l^{(1)}, v_{l+1}^{(1)})} (x-t)_+ \, d\nu_1(t) \\
&\quad = \int_{[v_l^{(1)}, v_{l+1}^{(1)})} \left(\big(x - v_l^{(1)}\big) \cdot \frac{v_{l+1}^{(1)} - t}{v_{l+1}^{(1)} - v_l^{(1)}} + \big(x - v_{l+1}^{(1)}\big) \cdot \frac{t - v_l^{(1)}}{v_{l+1}^{(1)} - v_l^{(1)}} \right) d\nu_1(t) \\
&\quad = \big(x - v_l^{(1)}\big) \cdot \widetilde{\mu}_l\big(\{v_l^{(1)}\}\big) + \big(x - v_{l+1}^{(1)}\big) \cdot \widetilde{\mu}_l\big(\{v_{l+1}^{(1)}\}\big) = \int_{[0,1]} (x-t)_+ \, d\widetilde{\mu}_l(t).
\end{aligned}$$

We also have that

$$\begin{aligned}
&|\widetilde{\mu}_l|([0,1]) = |\widetilde{\mu}_l|\big(\{v_l^{(1)}\}\big) + |\widetilde{\mu}_l|\big(\{v_{l+1}^{(1)}\}\big) \\
&\quad \leq \int_{[v_l^{(1)}, v_{l+1}^{(1)})} \frac{v_{l+1}^{(1)} - t}{v_{l+1}^{(1)} - v_l^{(1)}}\, d|\nu_1|(t) + \int_{[v_l^{(1)}, v_{l+1}^{(1)})} \frac{t - v_l^{(1)}}{v_{l+1}^{(1)} - v_l^{(1)}}\, d|\nu_1|(t) = |\nu_1|\big(\big[v_l^{(1)}, v_{l+1}^{(1)}\big)\big),
\end{aligned}$$

and that

$$|\widetilde{\mu}_0|((0,1]) = |\widetilde{\mu}_0|\big(\{v_1^{(1)}\}\big) \leq \int_{[0, v_1^{(1)})} \frac{t}{v_1^{(1)}}\, d|\nu_1|(t) \leq |\nu_1|\big(\big(0, v_1^{(1)}\big)\big).$$

Let $\widetilde{\mu}$ be the signed measure on $[0,1]$ defined by

$$\widetilde{\mu} = \sum_{l=0}^{n_1 - 1} \widetilde{\mu}_l$$

and let μ be the signed measure on $[0,1)$ defined by $\mu(E) = \widetilde{\mu}(E)$ for all Borel sets $E \subseteq [0,1)$. Note that μ is a discrete signed measure supported on $\mathcal{V}_1 \cap [0,1)$. For each $v_i^{(1)} \in \mathcal{V}_1$,

$$\begin{aligned} f_{c,\{\nu_1\}}\big(v_i^{(1)}\big) &= c + \int_{[0,1)} \big(v_i^{(1)} - t\big)_+ \, d\nu_1(t) = c + \sum_{l=0}^{n_1-1} \int_{[v_l^{(1)}, v_{l+1}^{(1)})} \big(v_i^{(1)} - t\big)_+ \, d\nu_1(t) \\ &= c + \sum_{l=0}^{i-1} \int_{[v_l^{(1)}, v_{l+1}^{(1)})} \big(v_i^{(1)} - t\big)_+ \, d\nu_1(t) = c + \int_{[0,1]} \big(v_i^{(1)} - t\big)_+ \, d\widetilde{\mu}(t) \\ &= c + \int_{[0,1)} \big(v_i^{(1)} - t\big)_+ \, d\mu(t) = f_{c,\{\mu\}}\big(v_i^{(1)}\big), \end{aligned}$$

which means that $f_{c,\{\mu\}}$ coincides with $f_{c,\{\nu_1\}}$ at every point of $\mathcal{V}_1$. Also,

$$\begin{aligned} |\mu|((0,1)) \le |\widetilde{\mu}|((0,1]) &\le \sum_{l=0}^{n_1-1} |\widetilde{\mu}_l|((0,1]) = |\widetilde{\mu}_0|((0,1]) + \sum_{l=1}^{n_1-1} |\widetilde{\mu}_l|([0,1]) \\ &\le |\nu_1|\big(\big(0, v_1^{(1)}\big)\big) + \sum_{l=1}^{n_1-1} |\nu_1|\big(\big[v_l^{(1)}, v_{l+1}^{(1)}\big)\big) = |\nu_1|((0,1)), \end{aligned}$$

from which we can conclude that

$$V_{\mathrm{mars}}(f_{c,\{\mu\}}) \le V_{\mathrm{mars}}(f_{c,\{\nu_1\}}).$$

Now, we consider general d and s. For $S \subseteq [d]$ with $0 < |S| \le s$ and $\mathbf{l} = (l_j, j \in S) \in \prod_{j \in S}\{0, \dots, n_j - 1\}$, let $\widetilde{\mu}_{S,\mathbf{l}}$ be the discrete signed measure on $[0,1]^{|S|}$ that is supported on $\prod_{j \in S}\{v_{l_j}^{(j)}, v_{l_j+1}^{(j)}\}$ and has masses

$$\widetilde{\mu}_{S,\mathbf{l}}\big(\big\{\big(v_{l_j+\delta_j}^{(j)}, j \in S\big)\big\}\big) = \int_{R_{\mathbf{l}}^S} \prod_{j \in S} \left[\left(\frac{v_{l_j+1}^{(j)} - t_j}{v_{l_j+1}^{(j)} - v_{l_j}^{(j)}} \right)^{1-\delta_j} \left(\frac{t_j - v_{l_j}^{(j)}}{v_{l_j+1}^{(j)} - v_{l_j}^{(j)}} \right)^{\delta_j} \right] d\nu_S(t_j, j \in S)$$

for each $(\delta_j, j \in S) \in \{0,1\}^{|S|}$, where $R_{\mathbf{l}}^S := \prod_{j \in S}[v_{l_j}^{(j)}, v_{l_j+1}^{(j)})$. Then, if $x_j \ge v_{l_j+1}^{(j)}$ for all $j \in S$, it follows that

$$\begin{aligned} &\int_{R_{\mathbf{l}}^S} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S) \\ &= \int_{R_{\mathbf{l}}^S} \prod_{j \in S} \left[\big(x_j - v_{l_j}^{(j)}\big) \cdot \frac{v_{l_j+1}^{(j)} - t_j}{v_{l_j+1}^{(j)} - v_{l_j}^{(j)}} + \big(x_j - v_{l_j+1}^{(j)}\big) \cdot \frac{t_j - v_{l_j}^{(j)}}{v_{l_j+1}^{(j)} - v_{l_j}^{(j)}} \right] d\nu_S(t_j, j \in S) \\ &= \int_{R_{\mathbf{l}}^S} \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} \prod_{j \in S} \left[\big(x_j - v_{l_j+\delta_j}^{(j)}\big) \left(\frac{v_{l_j+1}^{(j)} - t_j}{v_{l_j+1}^{(j)} - v_{l_j}^{(j)}} \right)^{1-\delta_j} \left(\frac{t_j - v_{l_j}^{(j)}}{v_{l_j+1}^{(j)} - v_{l_j}^{(j)}} \right)^{\delta_j} \right] d\nu_S(t_j, j \in S) \end{aligned}$$

$$= \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} \prod_{j \in S} \left(x_j - v^{(j)}_{l_j+\delta_j}\right) \cdot \int_{R^S_{\mathbf{l}}} \prod_{j \in S} \left[\left(\frac{v^{(j)}_{l_j+1} - t_j}{v^{(j)}_{l_j+1} - v^{(j)}_{l_j}} \right)^{1-\delta_j} \left(\frac{t_j - v^{(j)}_{l_j}}{v^{(j)}_{l_j+1} - v^{(j)}_{l_j}} \right)^{\delta_j} \right] d\nu_S(t_j, j \in S)$$
$$= \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} \prod_{j \in S} \left(x_j - v^{(j)}_{l_j+\delta_j}\right) \cdot \widetilde{\mu}_{S,\mathbf{l}}\left(\left\{\left(v^{(j)}_{l_j+\delta_j}, j \in S\right)\right\}\right)$$
$$= \int_{[0,1]^{|S|}} \prod_{j \in S} (x_j - t_j)_+ \, d\widetilde{\mu}_{S,\mathbf{l}}(t_j, j \in S).$$

Also,

$$\begin{aligned} |\widetilde{\mu}_{S,\mathbf{l}}|\left([0,1]^{|S|}\right) &= \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} \left| \int_{R^S_{\mathbf{l}}} \prod_{j \in S} \left[\left(\frac{v^{(j)}_{l_j+1} - t_j}{v^{(j)}_{l_j+1} - v^{(j)}_{l_j}} \right)^{1-\delta_j} \left(\frac{t_j - v^{(j)}_{l_j}}{v^{(j)}_{l_j+1} - v^{(j)}_{l_j}} \right)^{\delta_j} \right] d\nu_S(t_j, j \in S) \right| \\ &\le \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} \int_{R^S_{\mathbf{l}}} \prod_{j \in S} \left[\left(\frac{v^{(j)}_{l_j+1} - t_j}{v^{(j)}_{l_j+1} - v^{(j)}_{l_j}} \right)^{1-\delta_j} \left(\frac{t_j - v^{(j)}_{l_j}}{v^{(j)}_{l_j+1} - v^{(j)}_{l_j}} \right)^{\delta_j} \right] d|\nu_S|(t_j, j \in S) \\ &= \int_{R^S_{\mathbf{l}}} \prod_{j \in S} \left(\frac{t_j - v^{(j)}_{l_j}}{v^{(j)}_{l_j+1} - v^{(j)}_{l_j}} + \frac{v^{(j)}_{l_j+1} - t_j}{v^{(j)}_{l_j+1} - v^{(j)}_{l_j}} \right) d|\nu_S|(t_j, j \in S) = |\nu_S|\left(R^S_{\mathbf{l}}\right), \end{aligned}$$

and

$$|\widetilde{\mu}_{S,\mathbf{0}}|\left([0,1]^{|S|} \setminus \{\mathbf{0}\}\right) = \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|} \setminus \{\mathbf{0}\}} \left| \int_{R^S_{\mathbf{0}}} \prod_{j \in S} \left[\left(\frac{v^{(j)}_1 - t_j}{v^{(j)}_1 - 0} \right)^{1-\delta_j} \left(\frac{t_j - 0}{v^{(j)}_1 - 0} \right)^{\delta_j} \right] d\nu_S(t_j, j \in S) \right|$$
$$= \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|} \setminus \{\mathbf{0}\}} \left| \int_{R^S_{\mathbf{0}} \setminus \{\mathbf{0}\}} \prod_{j \in S} \left[\left(\frac{v^{(j)}_1 - t_j}{v^{(j)}_1 - 0} \right)^{1-\delta_j} \left(\frac{t_j - 0}{v^{(j)}_1 - 0} \right)^{\delta_j} \right] d\nu_S(t_j, j \in S) \right|$$
$$\le |\nu_S|\left(R^S_{\mathbf{0}} \setminus \{\mathbf{0}\}\right).$$

Now, let $\widetilde{\mu}_S$ be the signed measure on $[0,1]^{|S|}$ defined by

$$\widetilde{\mu}_S = \sum_{\mathbf{l} \in \prod_{j \in S} \{0,\ldots,n_j-1\}} \widetilde{\mu}_{S,\mathbf{l}}$$

and let μ_S be the signed measure on $[0,1)^{|S|}$ defined by

$$\mu_S(E) = \widetilde{\mu}_S(E)$$

for all Borel sets $E \subseteq [0,1)^{|S|}$. Note that μ_S is a discrete signed measure supported on $(\prod_{j \in S} \mathcal{V}_j) \cap [0,1)^{|S|}$. Then, for each $(v^{(1)}_{i_1}, \ldots, v^{(d)}_{i_d}) \in \prod_{j=1}^d \mathcal{V}_j$,

$$f_{c,\{\nu_S\}}\left(v^{(1)}_{i_1}, \ldots, v^{(d)}_{i_d}\right) = c + \sum_{0 < |S| \le s} \int_{[0,1)^{|S|}} \prod_{j \in S} \left(v^{(j)}_{i_j} - t_j\right)_+ d\nu_S(t_j, j \in S)$$

$$= c + \sum_{0<|S|\le s} \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\dots,n_j-1\}} \int_{R^S_{\mathbf{l}}} \prod_{j\in S} \big(v^{(j)}_{i_j} - t_j\big)_+ \, d\nu_S(t_j, j\in S)$$
$$= c + \sum_{0<|S|\le s} \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\dots,i_j-1\}} \int_{R^S_{\mathbf{l}}} \prod_{j\in S} \big(v^{(j)}_{i_j} - t_j\big)_+ \, d\nu_S(t_j, j\in S)$$
$$= c + \sum_{0<|S|\le s} \int_{[0,1]^{|S|}} \prod_{j\in S} \big(v^{(j)}_{i_j} - t_j\big)_+ \, d\widetilde{\mu}_S(t_j, j\in S)$$
$$= c + \sum_{0<|S|\le s} \int_{[0,1)^{|S|}} \prod_{j\in S} \big(v^{(j)}_{i_j} - t_j\big)_+ \, d\mu_S(t_j, j\in S) = f_{c,\{\mu_S\}}\big(v^{(1)}_{i_1}, \dots, v^{(d)}_{i_d}\big),$$

which means that $f_{c,\{\mu_S\}}$ agrees with $f_{c,\{\nu_S\}}$ at every point of $\prod_{j=1}^d \mathcal{V}_j$. Also, for every $S \subseteq [d]$ with $0 < |S| \le s$,

$$|\mu_S|\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big) \le |\widetilde{\mu}_S|\big([0,1]^{|S|} \setminus \{\mathbf{0}\}\big) \le \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\dots,n_j-1\}} |\widetilde{\mu}_{S,\mathbf{l}}|\big([0,1]^{|S|} \setminus \{\mathbf{0}\}\big)$$
$$= |\widetilde{\mu}_{S,\mathbf{0}}|\big([0,1]^{|S|} \setminus \{\mathbf{0}\}\big) + \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\dots,n_j-1\}\setminus\{\mathbf{0}\}} |\widetilde{\mu}_{S,\mathbf{l}}|\big([0,1]^{|S|}\big)$$
$$\le |\nu_S|\big(R^S_{\mathbf{0}} \setminus \{\mathbf{0}\}\big) + \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\dots,n_j-1\}\setminus\{\mathbf{0}\}} |\nu_S|\big(R^S_{\mathbf{l}}\big) = |\nu_S|\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big).$$

From this result, we can derive that

$$V_{\text{mars}}(f_{c,\{\mu_S\}}) \le V_{\text{mars}}(f_{c,\{\nu_S\}}).$$

□

C.6 Proof of Proposition 3

Proof of Proposition 3. Assume c is a real number and $\{\nu_S\}$ is a collection of finite signed measures, where ν_S is defined on $[0,1)^{|S|}$ for each $S \subseteq [d]$ with $0 < |S| \le s$. Recall that under this assumption,

$$f_{c,\{\nu_S\}}(x_1, \dots, x_d) = c + \sum_{0<|S|\le s} \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\dots,n_j-1\}} \nu_S\big(\big\{\big(v^{(j)}_{l_j}, j\in S\big)\big\}\big) \cdot \prod_{j\in S} \big(x_j - v^{(j)}_{l_j}\big)_+$$

and

$$V_{\text{mars}}(f_{c,\{\nu_S\}}) = \sum_{0<|S|\le s} \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\dots,n_j-1\}\setminus\{\mathbf{0}\}} \Big|\nu_S\big(\big\{\big(v^{(j)}_{l_j}, j\in S\big)\big\}\big)\Big|.$$

Now, let $\boldsymbol{\gamma}$ be the $|J|$-dimensional vector indexed by $(S, \mathbf{l}) \in J$ for which

$$\gamma_{S,\mathbf{l}} = \nu_S\big(\big\{\big(v^{(j)}_{l_j}, j\in S\big)\big\}\big)$$

for $(S, \mathbf{l}) \in J$. Then, it follows that

$$\big(f_{c,\{\nu_S\}}(\mathbf{x}^{(i)}), i = 1, \dots, n\big) = c\mathbf{1} + \mathbf{M}\boldsymbol{\gamma}$$

and

$$V_{\mathrm{mars}}(f_{c,\{\nu_S\}}) = \sum_{\substack{(S,\mathbf{l}) \in J \\ \mathbf{l} \neq \mathbf{0}}} |\gamma_{S,\mathbf{l}}|.$$

Thus, under the assumption, our estimation problem reduces to the finite-dimensional lasso problem (3.16). Recall that by Lemma 2, a solution to (3.5) obtained under the assumption is still a solution to the same problem without the assumption. Therefore, for a solution $(\hat{c}, \hat{\boldsymbol{\gamma}}_{n,V}^{d,s})$ to the problem (3.16), the function f on $[0,1]^d$ of the form

$$f(x_1, \dots, x_d) = \hat{c} + \sum_{(S,\mathbf{l}) \in J} \big(\hat{\gamma}_{n,V}^{d,s}\big)_{S,\mathbf{l}} \cdot \prod_{j \in S} \big(x_j - v_{l_j}^{(j)}\big)_+$$

is indeed a solution to the original problem (3.5). Moreover, since the set

$$\Bigg\{c\mathbf{1} + \mathbf{M}\boldsymbol{\gamma} : c \in \mathbb{R}, \boldsymbol{\gamma} \in \mathbb{R}^{|J|}, \text{ and } \sum_{\substack{(S,\mathbf{l}) \in J \\ \mathbf{l} \neq \mathbf{0}}} |\gamma_{S,\mathbf{l}}| \leq V\Bigg\}$$

is closed and convex, even though there can exist multiple solutions to the problem (3.16), they all should share the same value of $c\mathbf{1} + \mathbf{M}\boldsymbol{\gamma}$, which is indeed the projection of $\mathbf{y}$ onto the closed convex set. For these reasons, Lemma 2 also implies that for every solution $\hat{f}_{n,V}^{d,s}$ to the problem (3.5), the values of $\hat{f}_{n,V}^{d,s}$ at the design points should be

$$\hat{f}_{n,V}^{d,s}(\mathbf{x}^{(i)}) = \hat{c} + \big(\mathbf{M}\hat{\boldsymbol{\gamma}}_{n,V}^{d,s}\big)_i = \hat{c} + \sum_{(S,\mathbf{l}) \in J} \big(\hat{\gamma}_{n,V}^{d,s}\big)_{S,\mathbf{l}} \cdot \prod_{j \in S} \big(x_j^{(i)} - v_{l_j}^{(j)}\big)_+ \quad \text{for every } i = 1, \dots, n.$$

□

C.7 Proof of Theorem 2

We exploit the following general result for nonparametric least squares estimators. This result is a straightforward generalization of [58, Theorem 9.1] to the misspecified case, in which the true underlying function may not belong to the function class over which estimators are searched. Although we only need to consider the well-specified case for Theorem 2, here we state a generalized version for future purposes. We will use this theorem again for our proof of Theorem 4.

Theorem 20. *Suppose that the data $(\mathbf{x}^{(1)}, y_1), \dots, (\mathbf{x}^{(n)}, y_n)$ are generated according to the model*

$$y_i = f^*(\mathbf{x}^{(i)}) + \xi_i \tag{C.7}$$

where ξ_i are independent sub-Gaussian errors with mean zero and with a sub-Gaussian parameter σ, and f^* is an unknown real-valued regression function defined on $[0,1]^d$. Let $\mathcal{F}$ be a collection of real-valued functions defined on $[0,1]^d$ and let f_0 be an element of $\mathcal{F}$. Also, let $\hat{f}$ be a least squares estimator of f^* over $\mathcal{F}$ defined by

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \Big\{ \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}^{(i)})\big)^2 \Big\}.$$

Assume that there exists a real-valued function Ψ defined on $\mathbb{R}_{>0}$ such that $t \mapsto \Psi(t)/t^2$ is monotonically decreasing on $\mathbb{R}_{>0}$ and

$$\Psi(t) \geq \max \Big\{ t, \int_0^t \sqrt{\log N\big(\epsilon, B_{\mathcal{F}}(t, f_0, \|\cdot\|_n), \|\cdot\|_n\big)} d\epsilon \Big\}, \tag{C.8}$$

where $B_{\mathcal{F}}(t, f_0, \|\cdot\|_n) := \{f \in \mathcal{F} : \|f - f_0\|_n \leq t\}$ and $N(\epsilon, \mathcal{G}, \|\cdot\|_n)$ is the ϵ-covering number of a function class $\mathcal{G}$ under the norm $\|\cdot\|_n$. Then, there exist universal positive constants c and C such that

$$\mathbb{P}(\|\hat{f} - f^*\|_n^2 - \|f_0 - f^*\|_n^2 > t^2) \leq C \exp\Big(-\frac{nt^2}{C\sigma^2}\Big) \quad \text{for all } t \geq t_n, \tag{C.9}$$

for every $t_n > 0$ satisfying

$$\sqrt{n} t_n^2 \geq c\sigma\Psi(t_n) \quad \text{and} \quad t_n \geq \|f_0 - f^*\|_n.$$

Remark 8. *The probability bound* (C.9) *in Theorem 20 directly leads to an upper bound on the risk $R_F(\hat{f}, f^*)$. First, we can rewrite the bound* (C.9) *as*

$$\mathbb{P}(\|\hat{f} - f^*\|_n^2 - \|f_0 - f^*\|_n^2 > t) \leq C \exp\Big(-\frac{nt}{C\sigma^2}\Big) \quad \text{for all } t \geq t_n^2.$$

If we integrate both sides of the above inequality from t_n^2 to infinity, we obtain that

$$\mathbb{E}[(\|\hat{f} - f^*\|_n^2 - \|f_0 - f^*\|_n^2 - t_n^2)_+] \leq C \int_{t_n^2}^{\infty} \exp\Big(-\frac{nt}{C\sigma^2}\Big) dt \leq \frac{C\sigma^2}{n}.$$

Therefore, using the inequality $x \leq (x-y)_+ + y$, we can derive that

$$R_F(\hat{f}, f^*) = \mathbb{E}\|\hat{f} - f^*\|_n^2 \leq \|f_0 - f^*\|_n^2 + \frac{C\sigma^2}{n} + t_n^2. \tag{C.10}$$

Remark 9 (Well-specified Case)**.** *If f^* belongs to the function class $\mathcal{F}$, then we can choose f_0 as f^*. In this case, the condition $t_n \geq \|f_0 - f^*\|_n$ is vacuous, and* (C.10) *becomes*

$$R_F(\hat{f}, f^*) \leq \frac{C\sigma^2}{n} + t_n^2.$$

We will use this result for our proof of Theorem 2.

Remark 10. *Here, we briefly discuss how [58, Theorem 9.1] can be generalized to the misspecified case. We describe what needs to be modified in the proof of [58, Theorem 9.1] to achieve the generalized result in Theorem 20.*

Fix $t \geq t_n$ and let

$$\mathcal{F}_j = \{f \in \mathcal{F} : 2^{2j-2}t^2 < \|f - f^*\|_n^2 - \|f_0 - f^*\|_n^2 \leq 2^{2j}t^2\}$$

for each positive integer j. It then follows that

$$\mathbb{P}(\|\hat{f} - f^*\|_n^2 - \|f_0 - f^*\|_n^2 > t^2) = \sum_{j=1}^{\infty} \mathbb{P}(\hat{f} \in \mathcal{F}_j), \tag{C.11}$$

and it suffices to bound $\mathbb{P}(\hat{f} \in \mathcal{F}_j)$ for each positive integer j. Observe that for every $f \in \mathcal{F}_j$,

$$\begin{aligned} 2^{2j}t^2 &\geq \|f - f^*\|_n^2 - \|f_0 - f^*\|_n^2 = \|f - f_0\|_n^2 + \frac{2}{n}\sum_{i=1}^{n}(f - f_0)(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)}) \\ &\geq \|f - f_0\|_n^2 - 2\|f - f_0\|_n \cdot \|f_0 - f^*\|_n, \end{aligned}$$

and thus,

$$\|f - f_0\|_n \leq \|f_0 - f^*\|_n + \sqrt{\|f_0 - f^*\|_n^2 + 2^{2j}t^2} \leq 2\|f_0 - f^*\|_n + 2^j t \leq 2^{j+1}t.$$

Also, since $\hat{f}$ is a least squares estimator over $\mathcal{F}$ and f_0 is an element of $\mathcal{F}$,

$$\sum_{i=1}^{n}\left(y_i - \hat{f}(\mathbf{x}^{(i)})\right)^2 \leq \sum_{i=1}^{n}\left(y_i - f_0(\mathbf{x}^{(i)})\right)^2,$$

from which one can easily derive that

$$\|\hat{f} - f^*\|_n^2 - \|f_0 - f^*\|_n^2 \leq \frac{2}{n}\sum_{i=1}^{n}\xi_i(\hat{f} - f_0)(\mathbf{x}^{(i)}).$$

For these reasons, for each positive integer j, we have

$$\begin{aligned} \mathbb{P}(\hat{f} \in \mathcal{F}_j) &\leq \mathbb{P}\left(\sup_{f \in \mathcal{F}_j}\frac{1}{n}\sum_{i=1}^{n}\xi_i(f - f_0)(\mathbf{x}^{(i)}) \geq 2^{2j-3}t^2\right) \\ &\leq \mathbb{P}\left(\sup_{\substack{f \in \mathcal{F} \\ \|f - f_0\|_n \leq 2^{j+1}t}}\frac{1}{n}\sum_{i=1}^{n}\xi_i(f - f_0)(\mathbf{x}^{(i)}) \geq 2^{2j-3}t^2\right). \end{aligned}$$

From this point, we can simply follow the proof of [58, Theorem 9.1]. By repeating their argument, we can show that

$$\mathbb{P}(\hat{f} \in \mathcal{F}_j) \leq 4\exp\left(-\frac{2^j n t^2}{C\sigma^2}\right)$$

for each positive integer j. Combining this result with (C.11)*, we can conclude that*

$$\mathbb{P}(\|\hat{f} - f^*\|_n^2 - \|f_0 - f^*\|_n^2 > t^2) \leq \sum_{j=1}^{\infty} 4 \exp\Big(- \frac{2^j n t^2}{C\sigma^2}\Big) \leq C \exp\Big(- \frac{nt^2}{C\sigma^2}\Big).$$

Proof of Theorem 2. Let $\mathcal{F}_{\text{disc}}(V)$ be the collection of all functions $f_{c,\{\nu_S\}} \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$ with $V_{\text{mars}}(f_{c,\{\nu_S\}}) \leq V$ where ν_S is supported on $(\prod_{j \in S} \mathcal{V}_j) \cap [0,1)^{|S|}$ for each $S \subseteq [d]$ with $0 < |S| \leq s$. By Lemma 2, there exist $\hat{f}_{\text{disc}}, f^*_{\text{disc}} \in \mathcal{F}_{\text{disc}}(V)$ such that

$$\hat{f}_{\text{disc}}(\mathbf{x}^{(i)}) = \hat{f}^{d,s}_{n,V}(\mathbf{x}^{(i)}) \quad \text{and} \quad f^*_{\text{disc}}(\mathbf{x}^{(i)}) = f^*(\mathbf{x}^{(i)})$$

for all $i = 1, \dots, n$. Observe that

- $y_i = f^*(\mathbf{x}^{(i)}) + \xi_i = f^*_{\text{disc}}(\mathbf{x}^{(i)}) + \xi_i$ for $i = 1, \dots, n$;
- $R_F(\hat{f}^{d,s}_{n,V}, f^*) = R_F(\hat{f}_{\text{disc}}, f^*_{\text{disc}})$.

Also, we have

$$\hat{f}_{\text{disc}} \in \operatorname*{argmin}_{f \in \mathcal{F}_{\text{disc}}(V)} \bigg\{ \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}^{(i)})\big)^2 \bigg\}.$$

Thus, once we bound the metric entropy of $B_{\mathcal{F}_{\text{disc}}(V)}(t, f^*_{\text{disc}}, \|\cdot\|_n)$ (under the norm $\|\cdot\|_n$) and its integral as in (C.8), we can use Theorem 20 (and Remark 9) to obtain an upper bound on the risk $R_F(\hat{f}^{d,s}_{n,V}, f^*)$.

For $V > 0$ and $t > 0$, let

$$D(V,t) = \big\{ f \in \mathcal{F}_{\text{disc}}(V) : \|f\|_n \leq t \big\}.$$

Then, clearly,

$$B_{\mathcal{F}_{\text{disc}}(V)}(t, f^*_{\text{disc}}, \|\cdot\|_n) - \{f^*_{\text{disc}}\} \subseteq D(2V, t)$$

because

$$V_{\text{mars}}(f - f^*_{\text{disc}}) \leq V_{\text{mars}}(f) + V_{\text{mars}}(f^*_{\text{disc}}) \leq 2V$$

for every $f \in \mathcal{F}_{\text{disc}}(V)$. For two function classes $\mathcal{G}_1$ and $\mathcal{G}_2$, $\mathcal{G}_1 - \mathcal{G}_2$ indicates the class $\{g_1 - g_2 : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$. Also, the following lemma, which will be proved in Appendix C.15, provides an upper bound on the metric entropy of $D(V,t)$ (under the norm $\|\cdot\|_n$).

Lemma 12. *There exist positive constants $C_{\rho,s}$ depending on ρ and s and $C_{\rho,d}$ depending on ρ and d such that for every $V > 0$, $t > 0$, and $\epsilon > 0$,*

$$\begin{aligned}\log N(\epsilon, D(V,t), \|\cdot\|_n) &\leq 2^d \log\Big(2 + C_{\rho,s} \cdot \frac{V + \sqrt{n}t}{\epsilon}\Big) \\ &\quad + C_{\rho,d}\Big(2^{d+1} + 1 + \frac{2^{d+1}V}{\epsilon}\Big)^{1/2} \Big[\log\Big(2^{d+1} + 1 + \frac{2^{d+1}V}{\epsilon}\Big)\Big]^{3(2s-1)/4}.\end{aligned}$$

By Lemma 12 and the inequality $(x+y)^{1/2} \le x^{1/2} + y^{1/2}$, we can obtain the following upper bound on the square root of the metric entropy of $B_{\mathcal{F}_{\mathrm{disc}}(V)}(t, f^*_{\mathrm{disc}}, \|\cdot\|_n)$:

$$\sqrt{\log N\big(\epsilon, B_{\mathcal{F}_{\mathrm{disc}}(V)}(t, f^*_{\mathrm{disc}}, \|\cdot\|_n), \|\cdot\|_n\big)} \le 2^{d/2} \cdot \sqrt{\log\Big(2 + C_{\rho,s} \cdot \frac{2V + \sqrt{n}t}{\epsilon}\Big)} + C_{\rho,d}\Big(2^{d+1} + 1 + \frac{2^{d+2}V}{\epsilon}\Big)^{1/4}\Big[\log\Big(2^{d+1} + 1 + \frac{2^{d+2}V}{\epsilon}\Big)\Big]^{3(2s-1)/8}. \tag{C.12}$$

To integrate the right-hand side of (C.12), we construct the lemma below. We omit the proof of (C.13), which is basically the result of simple application of integration by parts, and our proof of (C.14) is presented in Appendix C.16.

Lemma 13. *For every $u > t$,*

$$\int_0^t \Big[\log\frac{u}{\epsilon}\Big]^{1/2} d\epsilon = \frac{t}{2\sqrt{\tau}}(1 + 2\tau), \tag{C.13}$$

where $\tau = \log(u/t)$. Also, for every $u > t$ and $k > 0$,

$$\int_0^t \Big(\frac{u}{\epsilon}\Big)^{1/4}\Big[\log\frac{u}{\epsilon}\Big]^k d\epsilon \le C_k u^{1/4} t^{3/4}(1 + \tau^k), \tag{C.14}$$

where C_k is a constant depending on k.

By (C.13),

$$\begin{aligned}\int_0^t \sqrt{\log\Big(2 + C_{\rho,s}\cdot\frac{2V+\sqrt{n}t}{\epsilon}\Big)}\, d\epsilon &\le \int_0^t \sqrt{\log\Big(\frac{2t + C_{\rho,s}(2V + \sqrt{n}t)}{\epsilon}\Big)}\, d\epsilon \\ &\le Ct\Big[1 + 2\log\Big(2 + C_{\rho,s}\Big(\sqrt{n} + \frac{2V}{t}\Big)\Big)\Big] \le C_{\rho,s} t\log n + Ct\log\Big(1 + \frac{2V}{\sqrt{n}t}\Big).\end{aligned}$$

Also, by (C.14) and the inequality $(x+y)^{1/4} \le x^{1/4} + y^{1/4}$,

$$\begin{aligned}&\int_0^t \Big(2^{d+1} + 1 + \frac{2^{d+2}V}{\epsilon}\Big)^{1/4}\Big[\log\Big(2^{d+1} + 1 + \frac{2^{d+2}V}{\epsilon}\Big)\Big]^{3(2s-1)/8} d\epsilon \\ &\le \int_0^t \Big(\frac{(2^{d+1}+1)t + 2^{d+2}V}{\epsilon}\Big)^{1/4}\Big[\log\Big(\frac{(2^{d+1}+1)t + 2^{d+2}V}{\epsilon}\Big)\Big]^{3(2s-1)/8} d\epsilon \\ &\le C_d\big[(2^{d+1}+1)t + 2^{d+2}V\big]^{1/4} t^{3/4}\Big[1 + \Big\{\log\Big(2^{d+1} + 1 + \frac{2^{d+2}V}{t}\Big)\Big\}^{3(2s-1)/8}\Big] \\ &\le C_d\big[\{(2^{d+1}+1)t\}^{1/4} + (2^{d+2}V)^{1/4}\big] t^{3/4}\Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8} \\ &\le C_d t\Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8} + C_d V^{1/4} t^{3/4}\Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8}.\end{aligned}$$

Combining these results, we can derive that

$$\int_0^t \sqrt{\log N\big(\epsilon, B_{\mathcal{F}_{\mathrm{disc}}(V)}(t, f^*_{\mathrm{disc}}, \|\cdot\|_n), \|\cdot\|_n\big)}\, d\epsilon \le C_{\rho,d} t \log n + C_d t \log\Big(1 + \frac{2V}{\sqrt{n}t}\Big) + C_{\rho,d} t \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8} + C_{\rho,d} V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8}.$$

Now, consider the function Ψ defined as the right-hand side of the above inequality, i.e.,

$$\Psi(t) = C_{\rho,d} t \log n + C_d t \log\Big(1 + \frac{2V}{\sqrt{n}t}\Big) + C_{\rho,d} t \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8} + C_{\rho,d} V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8}$$

for $t > 0$. Then, $t \mapsto \Psi(t)/t^2$ is monotonically decreasing and

$$\Psi(t) \ge \max\bigg(t, \int_0^t \sqrt{\log N\big(\epsilon, B_{\mathcal{F}_{\mathrm{disc}}(V)}(t, f^*_{\mathrm{disc}}, \|\cdot\|_n), \|\cdot\|_n\big)}\, d\epsilon\bigg),$$

by definition. Also, note that when $t \ge \sigma/\sqrt{n}$, we have

$$\begin{aligned} C_{\rho,d} t \log n &\le \frac{\sqrt{n}t^2}{4c\sigma} \quad \text{if } \ t \ge C_{\rho,d} \frac{\sigma \log n}{\sqrt{n}}, \\ C_d t \log\Big(1 + \frac{2V}{\sqrt{n}t}\Big) &\le \frac{\sqrt{n}t^2}{4c\sigma} \quad \text{if } \ t \ge C_d \frac{\sigma}{\sqrt{n}} \log\Big(1 + \frac{2V}{\sigma}\Big), \\ C_{\rho,d} t \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8} &\le \frac{\sqrt{n}t^2}{4c\sigma} \quad \text{if } \ t \ge C_{\rho,d} \frac{\sigma}{\sqrt{n}} \Big[\log\Big(2 + \frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/8}, \end{aligned}$$

and

$$C_{\rho,d} V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8} \le \frac{\sqrt{n}t^2}{4c\sigma} \quad \text{if } \ t \ge C_{\rho,d} \frac{\sigma^{4/5} V^{1/5}}{n^{2/5}} \Big[\log\Big(2 + \frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/10}.$$

Therefore, if we let

$$t_n = \max\bigg(\frac{\sigma}{\sqrt{n}}, C_{\rho,d} \frac{\sigma \log n}{\sqrt{n}}, C_d \frac{\sigma}{\sqrt{n}} \log\Big(1 + \frac{2V}{\sigma}\Big), C_{\rho,d} \frac{\sigma}{\sqrt{n}} \Big[\log\Big(2 + \frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/8}, C_{\rho,d} \frac{\sigma^{4/5} V^{1/5}}{n^{2/5}} \Big[\log\Big(2 + \frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/10}\bigg),$$

then the inequality

$$\sqrt{n} t_n^2 \ge c\sigma \Psi(t_n)$$

holds. By Theorem 20 (and Remark 9), it follows that

$$
\begin{aligned}
R_F\big(\hat{f}^{d,s}_{n,V}, f^*\big) &\le \frac{C\sigma^2}{n} + t_n^2 \\
&\le C_{\rho,d}\Big(\frac{\sigma^2 V^{1/2}}{n}\Big)^{4/5}\Big[\log\Big(2+\frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/5} + C_{\rho,d}\frac{\sigma^2}{n}\Big[\log\Big(2+\frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/4} \\
&\quad + C_d\frac{\sigma^2}{n}\Big[\log\Big(1+\frac{2V}{\sigma}\Big)\Big]^2 + C_{\rho,d}\frac{\sigma^2}{n}[\log n]^2.
\end{aligned}
\tag{C.15}
$$

We can further observe that the second term and the third term on the right-hand side of (C.15) can be absorbed into the other terms. Note that there exist some positive constants C_s and C such that

$$
[\log(2+x)]^{3(2s-1)/4} \le x^{2/5} \ \text{ for } x \ge C_s \quad \text{and} \quad [\log(1+2x)]^2 \le x^{2/5} \ \text{ for } x \ge C.
$$

Hence, the second term can be bounded by

$$
C_{\rho,d}\frac{\sigma^2}{n}\Big[\log\Big(2+\frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/4} \le C_{\rho,d}\frac{\sigma^2}{n} + C_{\rho,d}\frac{\sigma^2}{n}\Big(\frac{Vn^{1/2}}{\sigma}\Big)^{2/5} = C_{\rho,d}\frac{\sigma^2}{n} + C_{\rho,d}\Big(\frac{\sigma^2V^{1/2}}{n}\Big)^{4/5},
$$

and the third term can be bounded by

$$
\begin{aligned}
C_d\frac{\sigma^2}{n}\Big[\log\Big(1+\frac{2V}{\sigma}\Big)\Big]^2 &\le C_d\frac{\sigma^2}{n}\Big[\log\Big(1+\frac{2Vn^{1/2}}{\sigma}\Big)\Big]^2 \\
&\le C_d\frac{\sigma^2}{n} + C_d\frac{\sigma^2}{n}\Big(\frac{Vn^{1/2}}{\sigma}\Big)^{2/5} = C_d\frac{\sigma^2}{n} + C_d\Big(\frac{\sigma^2V^{1/2}}{n}\Big)^{4/5}.
\end{aligned}
$$

Consequently, removing both terms from (C.15), we can derive that

$$
R_F\big(\hat{f}^{d,s}_{n,V}, f^*\big) \le C_{\rho,d}\Big(\frac{\sigma^2V^{1/2}}{n}\Big)^{4/5}\Big[\log\Big(2+\frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/5} + C_{\rho,d}\frac{\sigma^2}{n}[\log n]^2.
$$

□

C.8 Proof of Theorem 3

We first briefly recall what Brownian sheet is and introduce integrated Brownian sheet. An m-dimensional Brownian sheet $B_m = (B_m(\mathbf{t},\cdot) : \mathbf{t} \in [0,1]^m)$ is a centered Gaussian process on $[0,1]^m$ whose covariance is given by

$$
\mathbb{E}\big[B_m(\mathbf{t},\cdot)B_m(\mathbf{s},\cdot)\big] = \prod_{j=1}^{m} \min(t_j, s_j).
$$

Let $\{\psi_{k_1}\otimes\cdots\otimes\psi_{k_m} : k_1,\dots,k_m \in \mathbb{N}\}$ be an orthonormal basis of $L^2([0,1]^m)$. Here $\psi_{k_1}\otimes\cdots\otimes\psi_{k_m}$ indicates the tensor product of the functions $\psi_{k_1},\dots,\psi_{k_m}$, which is defined by

$$
(\psi_{k_1}\otimes\cdots\otimes\psi_{k_m})(\mathbf{x}) = \psi_{k_1}(x_1)\dots\psi_{k_m}(x_m)
$$

for $\mathbf{x} = (x_1, \ldots, x_m) \in [0,1]^m$. Then, it is well-known that the Brownian sheet B_m can be represented by

$$B_m(\cdot, \omega) = \sum_{k_1,\ldots,k_m \in \mathbb{N}} \xi_{k_1,\ldots,k_m}(\omega) \cdot \mathcal{T}_m(\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\cdot),$$

where $\xi_{k_1,\ldots,k_m}$ are independent standard normal random variables, $\mathcal{T}_m : L^2([0,1]^m) \to C([0,1]^m)$ is the m-dimensional integral operator defined by

$$\mathcal{T}_m f(\mathbf{t}) = \int_{\prod_{j=1}^m [0,t_j]} f(\mathbf{x})\, d\mathbf{x} \quad \text{for } \mathbf{t} = (t_1, \ldots, t_m) \in [0,1]^m,$$

and the series converges a.s. in $C([0,1]^m)$ (see, e.g., [60, Section 5]). An m-dimensional integrated Brownian sheet $X_m = (X_m(\mathbf{t}, \cdot) : \mathbf{t} \in [0,1]^m)$ is then simply defined as

$$X_m(\cdot, \omega) = \mathcal{T}_m\big(B_m(\cdot, \omega)\big).$$

Since $\mathcal{T}_m$ is a continuous linear operator on $C([0,1]^m)$, it follows that

$$\begin{aligned} X_m(\cdot, \omega) &= \mathcal{T}_m\Bigg(\sum_{k_1,\ldots,k_m \in \mathbb{N}} \xi_{k_1,\ldots,k_m}(\omega) \cdot \mathcal{T}_m(\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\cdot)\Bigg) \\ &= \sum_{k_1,\ldots,k_m \in \mathbb{N}} \xi_{k_1,\ldots,k_m}(\omega) \cdot \mathcal{T}_m^2(\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\cdot) \end{aligned}$$

almost surely. We will exploit this series expansion of the integrated Brownian sheet X_m in our proof of Theorem 3.

We also utilize the following two results for our proof of Theorem 3. The first result is [94, Theorem 1.2], which is restated as Theorem 21 below. This result implies that upper bounds on the metric entropy of $\mathcal{D}_m$ can be obtained from appropriate lower bounds on the small ball probability of the integrated Brownian sheet X_m. The second result is [30, Theorem 1.2], which is restated as Theorem 22 below. This result helps us connect the small ball probability of the integrated Brownian sheet X_m to that of the Brownian sheet B_m whose lower bounds are quite well-studied.

Recall that a random element X in a Banach space E is said to be centered Gaussian if $\langle A, X\rangle := A(X)$ is normally distributed and has mean zero for all $A \in E^*$, where E^* is the dual space of E. Also, its covariance operator $\Sigma : E^* \to E$ is defined by

$$\langle A, \Sigma B\rangle = \int_E \langle A, x\rangle \langle B, x\rangle \, d\mu(x) = \mathbb{E}[\langle A, X\rangle\langle B, X\rangle]$$

for all $A, B \in E^*$, where μ is the measure induced on E by X (see, e.g., [96, Section 8]). Moreover, recall that for the map $\Phi : E^* \to E$ defined by the Bochner integral

$$\Phi(A) = \int_E \langle A, x\rangle x \, d\mu(x),$$

the reproducing kernel Hilbert space of X in E is defined as the completion of $\Phi(E^*)$ with respect to the inner product

$$\langle \Phi(A), \Phi(B) \rangle := \int_E \langle A, x \rangle \langle B, x \rangle \, d\mu(x) = \mathbb{E}[\langle A, X \rangle \langle B, X \rangle]$$

(see, e.g., [151, Section 2]).

Theorem 21 (Theorem 1.2 of [94])**.** *Let $(E, \|\cdot\|)$ be a separable Banach space and X be a centered Gaussian element in E. Also, let K_X be the unit ball of the reproducing kernel Hilbert space of X in E. For $\alpha > 0$ and $\beta \in \mathbb{R}$, if there exists some constant $c > 0$ such that*

$$\log \mathbb{P}(\|X\| \le \epsilon) \ge -c\epsilon^{-\alpha} \Big[\log \frac{1}{\epsilon} \Big]^{\beta}$$

for sufficiently small $\epsilon > 0$, then there exists some constant $C > 0$ such that

$$\log N(\epsilon, K_X, \|\cdot\|) \le C\epsilon^{-2\alpha/(2+\alpha)} \Big[\log \frac{1}{\epsilon} \Big]^{2\beta/(2+\alpha)}$$

for sufficiently small $\epsilon > 0$.

Theorem 22 (Theorem 1.2 of [30])**.** *Let $(E, \|\cdot\|)$ be a separable Banach space and $(H, \|\cdot\|_H)$ be a Hilbert space. Also, let Y be a centered Gaussian element in H. Then, for every linear operator $L : H \to E$ and a centered Gaussian element X in E with the covariance operator LL^*, where $L^* : E^* \to H$ is the adjoint operator of L,*

$$\mathbb{P}(\|LY\| \le \epsilon) \ge \mathbb{P}(\|X\| \le \lambda\epsilon) \cdot \mathbb{E}\Big[\exp\Big(-\frac{\lambda^2}{2} \|Y\|_H^2 \Big) \Big]$$

for all $\lambda > 0$ and $\epsilon > 0$.

In our proof of Theorem 3, we use Theorem 21 with $E = C([0,1]^m)$ and $X = X_m$ and Theorem 22 with $E = C([0,1]^m)$, $H = L^2([0,1]^m)$, $L = \mathcal{T}_m$, and $X = Y = B_m$. Observe that the Brownian sheet B_m is a centered Gaussian element both in $L^2([0,1]^m)$ and $C([0,1]^m)$, and the integrated Brownian sheet X_m is a centered Gaussian element in $C([0,1]^m)$. It is also known that the covariance operator of the Brownian sheet B_m is $\mathcal{T}_m \mathcal{T}_m^*$ (see, e.g., [44, Section 3]). Moreover, one can check from the series expansion of X_m and the theorem below that the reproducing kernel Hilbert space H_{X_m} of X_m in $C([0,1]^m)$ and its unit ball K_{X_m} can be written as

$$H_{X_m} = \Big\{ \sum_{k_1, \dots, k_m \in \mathbb{N}} w_{k_1, \dots, k_m} \cdot \mathcal{T}_m^2(\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\cdot) : (w_{k_1, \dots, k_m}, k_1, \dots, k_m \in \mathbb{N}) \in \ell^2 \Big\}$$

and

$$K_{X_m} = \Big\{ \sum_{k_1, \dots, k_m \in \mathbb{N}} w_{k_1, \dots, k_m} \cdot \mathcal{T}_m^2(\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\cdot) : (w_{k_1, \dots, k_m}, k_1, \dots, k_m \in \mathbb{N}) \in B_{\ell^2} \Big\},$$

where B_{ℓ^2} is the unit ball in ℓ^2.

Theorem 23 (Theorem 4.2 of [151])**.** *Suppose that $\{h_k\}_{k\geq 1}$ is a sequence in a separable Banach space E such that if $w \in \ell^2$ and*

$$\sum_{k=1}^{\infty} w_k h_k = 0,$$

where the convergence is in E, then $w = 0$. Then, for a random element X in E defined by

$$X = \sum_{k=1}^{\infty} \xi_k h_k,$$

where ξ_k are independent standard normal random variables, and the series converges a.s. in E, the reproducing kernel Hilbert space of X in E is given by

$$H_X = \left\{ \sum_{k=1}^{\infty} w_k h_k : w \in \ell^2 \right\}$$

with the norm

$$\left\| \sum_{k=1}^{\infty} w_k h_k \right\| = \|w\|_2.$$

Proof of Theorem 3. For each $k \in \mathbb{N}$, let ϕ_k be the function on $[0,1]$ defined by $\phi_k(x) = \psi_k(1-x)$ for $x \in [0,1]$. Then, clearly, $\{\phi_{k_1} \otimes \cdots \otimes \phi_{k_m} : k_1, \ldots, k_m \in \mathbb{N}\}$ is also an orthonormal basis of $L^2([0,1]^m)$. Also, let

$$Z = \big\{ (z_{k_1,\ldots,k_m}, k_1, \ldots, k_m \in \mathbb{N}) : z_{k_1,\ldots,k_m} = \langle F, \phi_{k_1} \otimes \cdots \otimes \phi_{k_m} \rangle \\ \text{for } k_1, \ldots, k_m \in \mathbb{N}, \text{ where } F \in \mathcal{D}_m \big\}.$$

Note that Z is convex and symmetric with respect to the origin, because $\mathcal{D}_m$ is convex and symmetric with respect to the zero-function. Also, for $F \in \mathcal{D}_m$ of the form

$$F(x_1, \ldots, x_m) = \int (x_1 - t_1)_+ \cdots (x_m - t_m)_+ \, d\nu(\mathbf{t}) = \int_{\prod_{j=1}^m [0,x_j]} (x_1 - t_1) \cdots (x_m - t_m) \, d\nu(\mathbf{t}),$$

where ν is a signed measure on $[0,1]^m$ with total variation $|\nu|([0,1]^m) \leq 1$, we have

$$\begin{aligned} &\langle F, \phi_{k_1} \otimes \cdots \otimes \phi_{k_m} \rangle \\ &\quad = \int_{[0,1]^m} \int_{\prod_{j=1}^m [0,x_j]} (x_1 - t_1) \cdots (x_m - t_m) \, d\nu(\mathbf{t}) \cdot (\phi_{k_1} \otimes \cdots \otimes \phi_{k_m})(\mathbf{x}) \, d\mathbf{x} \qquad \text{(C.16)} \\ &\quad = \int_{[0,1]^m} \int_{t_1}^1 \cdots \int_{t_m}^1 (x_1 - t_1) \cdots (x_m - t_m) \cdot \phi_{k_1}(x_1) \cdots \phi_{k_m}(x_m) \, dx_m \cdots dx_1 \, d\nu(\mathbf{t}) \end{aligned}$$

for $k_1, \dots, k_m \in \mathbb{N}$. By Parseval's identity, Z is a subset of ℓ^2 and

$$\log N(\epsilon, \mathcal{D}_m, \|\cdot\|_2) = \log N(\epsilon, Z, \|\cdot\|_{\ell^2}). \tag{C.17}$$

Also, since Z is convex and symmetric, by the fundamental duality theorem of metric entropy (refer to, e.g., [3, Theorem 5]), there exist universal constants $c, C > 0$ such that

$$\log N(\epsilon, Z, \|\cdot\|_{\ell^2}) \le C \log N(c\epsilon, B_{\ell^2}, \|\cdot\|_{Z^\circ}), \tag{C.18}$$

where B_{ℓ^2} is the unit ball with respect to the ℓ^2 norm and $\|\cdot\|_{Z^\circ}$ is the gauge of

$$Z^\circ = \left\{ (w_{k_1,\dots,k_m}, k_1, \dots, k_m \in \mathbb{N}) \in \ell^2 : \sup_{z \in Z} \left| \sum_{k_1,\dots,k_m \in \mathbb{N}} w_{k_1,\dots,k_m} z_{k_1,\dots,k_m} \right| \le 1 \right\},$$

i.e.,

$$\|\mathbf{w}\|_{Z^\circ} = \inf\{r \ge 0 : \mathbf{w} \in rZ^\circ\}$$

for $\mathbf{w} = (w_{k_1,\dots,k_m}, k_1, \dots, k_m \in \mathbb{N}) \in \ell^2$. Due to (C.16), we can also write Z° as

$$\begin{aligned} Z^\circ = \Bigg\{ (w_{k_1,\dots,k_m}, k_1, \dots, k_m \in \mathbb{N}) \in \ell^2 : \sup_{\mathbf{t} \in [0,1]^m} \Bigg| \sum_{k_1,\dots,k_m \in \mathbb{N}} w_{k_1,\dots,k_m} \\ \cdot \int_{t_1}^1 \cdots \int_{t_m}^1 (x_1 - t_1) \cdots (x_m - t_m) \phi_{k_1}(x_1) \cdots \phi_{k_m}(x_m)\, dx_m \cdots dx_1 \Bigg| \le 1 \Bigg\}. \end{aligned} \tag{C.19}$$

We now relate the covering number $N(\epsilon, B_{\ell^2}, \|\cdot\|_{Z^\circ})$ to the small ball probability of the integrated Brownian sheet X_m. Recall that we can represent the integrated Brownian sheet X_m as

$$X_m(\cdot, \omega) = \sum_{k_1,\dots,k_m \in \mathbb{N}} \xi_{k_1,\dots,k_m}(\omega) \cdot \mathcal{T}_m^2(\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\cdot), \tag{C.20}$$

where the series converges a.s. in $C([0,1]^m)$. Observe that

$$\begin{aligned} \mathcal{T}_m^2(\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\mathbf{t}) &= \int_{\prod_{j=1}^m [0,t_j]} \int_{\prod_{j=1}^m [0,s_j]} (\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\mathbf{x})\, d\mathbf{x}\, d\mathbf{s} \\ &= \int_{\prod_{j=1}^m [0,t_j]} \int_{\prod_{j=1}^m [x_j,t_j]} (\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\mathbf{x})\, d\mathbf{s}\, d\mathbf{x} \\ &= \int_{\prod_{j=1}^m [0,t_j]} (t_1 - x_1) \cdots (t_m - x_m) \psi_{k_1}(x_1) \cdots \psi_{k_m}(x_m)\, d\mathbf{x} \\ &= \int_{s_1}^1 \cdots \int_{s_m}^1 (z_1 - s_1) \cdots (z_m - s_m) \phi_{k_1}(z_1) \cdots \phi_{k_m}(z_m)\, dz_m \cdots dz_1 \end{aligned} \tag{C.21}$$

for $\mathbf{t} \in [0,1]^m$, where $(s_1, \dots, s_m) = (1 - t_1, \dots, 1 - t_m)$. By (C.19) and (C.21),

$$\sup_{\mathbf{t} \in [0,1]^m} \left| \sum_{k_1,\dots,k_m \in \mathbb{N}} w_{k_1,\dots,k_m} \cdot \mathcal{T}_m^2(\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\mathbf{t}) \right|$$

$$= \sup_{\mathbf{s}\in[0,1]^m} \Bigg| \sum_{k_1,\dots,k_m\in\mathbb{N}} w_{k_1,\dots,k_m} \cdot \int_{s_1}^{1} \cdots \int_{s_m}^{1} (z_1 - s_1)\cdots(z_m - s_m)\phi_{k_1}(z_1)\cdots\phi_{k_m}(z_m)\, dz_m \cdots dz_1 \Bigg| = \|\mathbf{w}\|_{Z^\circ}$$

for $\mathbf{w} = (w_{k_1,\dots,k_m}, k_1, \dots, k_m \in \mathbb{N}) \in \ell^2$. Since the unit ball K_{X_m} of the reproducing kernel Hilbert space of X_m in $C([0,1]^m)$ is

$$K_{X_m} = \Big\{ \sum_{k_1,\dots,k_m\in\mathbb{N}} w_{k_1,\dots,k_m} \cdot \mathcal{T}_m^2(\psi_{k_1} \otimes \cdots \otimes \psi_{k_m})(\cdot) : (w_{k_1,\dots,k_m}, k_1, \dots, k_m \in \mathbb{N}) \in B_{\ell^2} \Big\},$$

it thus follows that

$$N(\epsilon, K_{X_m}, \|\cdot\|_\infty) = N(\epsilon, B_{\ell^2}, \|\cdot\|_{Z^\circ}).$$

Hence, for the integrated Brownian sheet X_m, we can restate Theorem 21 as follows: for $\alpha > 0$ and $\beta \in \mathbb{R}$,

$$\log \mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |X_m(\mathbf{t})| \le \epsilon \Big) \ge -c_m \epsilon^{-\alpha} \Big[\log \frac{1}{\epsilon} \Big]^{\beta} \text{ for all sufficiently small } \epsilon > 0 \tag{C.22}$$

$$\implies \log N(\epsilon, B_{\ell^2}, \|\cdot\|_{Z^\circ}) \le C_m \epsilon^{-2\alpha/(2+\alpha)} \Big[\log \frac{1}{\epsilon} \Big]^{2\beta/(2+\alpha)} \text{ for all sufficiently small } \epsilon > 0.$$

Combining (C.17), (C.18), and (C.22), we can see that it suffices to find lower bounds on $\log \mathbb{P}(\sup_{\mathbf{t}\in[0,1]^m} |X_m(\mathbf{t})| \le \epsilon)$ in order to find upper bounds of the metric entropy of $\mathcal{D}_m$.

We now turn to the small ball probability of X_m. Recall that the covariance operator of the Brownian sheet B_m is $\mathcal{T}_m\mathcal{T}_m^*$. By applying Theorem 22 to $E = C([0,1]^m)$, $H = L^2([0,1]^m)$, $L = \mathcal{T}_m$, and $X = Y = B_m$, we can derive that

$$\mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |X_m(\mathbf{t})| \le \epsilon \Big) \ge \mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |B_m(\mathbf{t})| \le \lambda\epsilon \Big) \cdot \mathbb{E}\Big[\exp\Big(-\frac{\lambda^2}{2} \int_{[0,1]^m} B_m^2(\mathbf{t})\, d\mathbf{t} \Big) \Big].$$

for all $\lambda > 0$ and $\epsilon > 0$. Substituting $\lambda = \epsilon^{-2/3}$ and using Markov's inequality, we obtain

$$\begin{aligned} \mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |X_m(\mathbf{t})| \le \epsilon \Big) &\ge \mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |B_m(\mathbf{t})| \le \epsilon^{1/3} \Big) \cdot \mathbb{E}\Big[\exp\Big(-\frac{\epsilon^{-4/3}}{2} \int_{[0,1]^m} B_m^2(\mathbf{t})\, d\mathbf{t} \Big) \Big] \\ &\ge \mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |B_m(\mathbf{t})| \le \epsilon^{1/3} \Big) \cdot \Big[\mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |B_m(\mathbf{t})| \le \epsilon^{1/3} \Big) \cdot \exp\Big(-\frac{\epsilon^{-2/3}}{2} \Big) \Big] \\ &\ge \Big[\mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |B_m(\mathbf{t})| \le \epsilon^{1/3} \Big) \Big]^2 \cdot \exp\Big(-\frac{\epsilon^{-2/3}}{2} \Big). \end{aligned}$$

In [44, Theorem 6], it was proved that

$$\log \mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |B_m(\mathbf{t})| \le \epsilon \Big) \ge -c_m \epsilon^{-2} \Big[\log \frac{1}{\epsilon} \Big]^{2m-1}$$

for sufficiently small $\epsilon > 0$. In particular, it is well-known that the logarithmic multiplicative factor of the right-hand side can be omitted when $m = 1$. Therefore, it follows that

$$\log \mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |X_m(\mathbf{t})| \le \epsilon\Big) \ge 2\log \mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^m} |B_m(\mathbf{t})| \le \epsilon^{1/3}\Big) - \frac{\epsilon^{-2/3}}{2}$$
$$\ge -c_m\epsilon^{-2/3}\Big[\log\frac{1}{\epsilon}\Big]^{2m-1} - \frac{\epsilon^{-2/3}}{2} \ge -c_m\epsilon^{-2/3}\Big[\log\frac{1}{\epsilon}\Big]^{2m-1}$$

for sufficiently small $\epsilon > 0$. This result together with (C.22) implies that

$$\log N(\epsilon, B_{\ell^2}, \|\cdot\|_{Z^\circ}) \le C_m\epsilon^{-1/2}\Big[\log\frac{1}{\epsilon}\Big]^{3(2m-1)/4}$$

for sufficiently small $\epsilon > 0$. Consequently, combining all the pieces, we can prove that

$$\log N(\epsilon, \mathcal{D}_m, \|\cdot\|_2) \le C_m\epsilon^{-1/2}\Big[\log\frac{1}{\epsilon}\Big]^{3(2m-1)/4}$$

for sufficiently small $\epsilon > 0$. Note that we can omit the logarithmic factor when $m = 1$. □

C.9 Proof of Theorem 4

Our proof of Theorem 4 uses the following lemma, which we prove in Appendix C.17. This lemma states that for every function $f \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$, we can find $f_{c,\{\mu_S\}}$ which is close to f and whose μ_S are discrete signed measures supported on the lattices generated by $\widetilde{\mathcal{V}}_j$ (recall (3.18)). We will use this lemma again for our proof of Theorem 7.

Lemma 14. *Suppose we are given a real number c and a collection of finite signed measures $\{\nu_S\}$ where ν_S is defined on $[0,1)^{|S|}$ for each $S \subseteq [d]$ with $0 < |S| \le s$. Then, there exists a collection of discrete signed measures $\{\mu_S\}$ where μ_S is supported on $(\prod_{j\in S}\widetilde{\mathcal{V}}_j) \cap [0,1)^{|S|}$ for each $S \subseteq [d]$ with $0 < |S| \le s$ such that*

- $V_{mars}(f_{c,\{\mu_S\}}) \le V_{mars}(f_{c,\{\nu_S\}})$;
- $\|f_{c,\{\mu_S\}} - f_{c,\{\nu_S\}}\|_\infty \le (2/N) \cdot V_{mars}(f_{c,\{\nu_S\}})$;
- $\|f_{c,\{\mu_S\}} - f_{c,\{\nu_S\}}\|_{p_0,2} \le C_d(B/N^3)^{1/2} \cdot V_{mars}(f_{c,\{\nu_S\}})$

for some positive constant C_d depending on d, where $N = \min_j N_j$.

Proof of Theorem 4. Recall the definition of $\mathcal{F}_{\text{disc}}(V)$ from our proof of Theorem 2. Also, let $\mathcal{F}_{\text{approx}}(V)$ be the collection of all functions $f_{c,\{\nu_S\}} \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$ with $V_{\text{mars}}(f_{c,\{\nu_S\}}) \le V$ where ν_S is supported on $(\prod_{j\in S}\widetilde{\mathcal{V}}_j) \cap [0,1)^{|S|}$ for each $S \subseteq [d]$ with $0 < |S| \le s$. Then, clearly,

$$\tilde{f}^{d,s}_{n,V} \in \operatorname*{argmin}_{f\in\mathcal{F}_{\text{approx}}(V)} \bigg\{\sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)})\big)^2\bigg\}.$$

Also, Lemma 14 guarantees that we have $f_0 \in \mathcal{F}_{\text{approx}}(V)$ for which

$$\|f_0 - f^*\|_n \le \|f_0 - f^*\|_\infty \le \frac{2V}{N}.$$

In order to apply Theorem 20 with $f_0 \in \mathcal{F}_{\text{approx}}(V)$, we need to bound the metric entropy of $B_{\mathcal{F}_{\text{approx}}(V)}(t, f_0, \|\cdot\|_n)$ (under the norm $\|\cdot\|_n$). By Lemma 2, there exists $g_0 \in \mathcal{F}_{\text{disc}}(V)$ such that

$$g_0(\mathbf{x}^{(i)}) = f_0(\mathbf{x}^{(i)})$$

for all $i = 1, \dots, n$. Also, by the same lemma and the fact that the norm $\|\cdot\|_n$ only depends on the values of functions at $\mathbf{x}^{(i)}$, we have

$$N\big(\epsilon, B_{\mathcal{F}_{\text{approx}}(V)}(t, f_0, \|\cdot\|_n), \|\cdot\|_n\big) \le N\big(\epsilon, B_{\mathcal{F}_{\text{disc}}(V)}(t, g_0, \|\cdot\|_n), \|\cdot\|_n\big)$$

for every $\epsilon > 0$. Hence, we can repeat our argument in the proof of Theorem 2 to bound $N(\epsilon, B_{\mathcal{F}_{\text{disc}}(V)}(t, g_0, \|\cdot\|_n), \|\cdot\|_n)$ and thereby show that the function Ψ defined by

$$\begin{aligned}\Psi(t) &= C_{\rho,d} t \log n + C_d t \log\Big(1 + \frac{2V}{\sqrt{n}t}\Big) \\ &\quad + C_{\rho,d} t \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8} + C_{\rho,d} V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{3(2s-1)/8}\end{aligned}$$

for $t > 0$ and

$$\begin{aligned}t_n = \max\Bigg(&\|f_0 - f^*\|_n, \frac{\sigma}{\sqrt{n}}, C_{\rho,d}\frac{\sigma \log n}{\sqrt{n}}, C_d \frac{\sigma}{\sqrt{n}} \log\Big(1 + \frac{2V}{\sigma}\Big), \\ &C_{\rho,d}\frac{\sigma}{\sqrt{n}}\Big[\log\Big(2 + \frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/8}, C_{\rho,d}\frac{\sigma^{4/5}V^{1/5}}{n^{2/5}}\Big[\log\Big(2 + \frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/10}\Bigg),\end{aligned}$$

satisfy the conditions in Theorem 20. As a result, Theorem 20 (and Remark 8) implies

$$\begin{aligned}R_F\big(\widetilde{f}^{d,s}_{n,V}, f^*\big) &\le \|f_0 - f^*\|_n^2 + \frac{C\sigma^2}{n} + t_n^2 \\ &\le \frac{8V^2}{N^2} + C_{\rho,d}\Big(\frac{\sigma^2 V^{1/2}}{n}\Big)^{4/5}\Big[\log\Big(2 + \frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/5} \\ &\quad + C_{\rho,d}\frac{\sigma^2}{n}\Big[\log\Big(2 + \frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/4} + C_d\frac{\sigma^2}{n}\Big[\log\Big(1 + \frac{2V}{\sigma}\Big)\Big]^2 + C_{\rho,d}\frac{\sigma^2}{n}[\log n]^2.\end{aligned}$$

As in the proof of Theorem 2, we can absorb the third and fourth terms into the other terms. Consequently, we can show that

$$R_F\big(\widetilde{f}^{d,s}_{n,V}, f^*\big) \le \frac{8V^2}{N^2} + C_{\rho,d}\Big(\frac{\sigma^2 V^{1/2}}{n}\Big)^{4/5}\Big[\log\Big(2 + \frac{Vn^{1/2}}{\sigma}\Big)\Big]^{3(2s-1)/5} + C_{\rho,d}\frac{\sigma^2}{n}[\log n]^2.$$

□

C.10 Proof of Theorem 5

The following theorem plays a central role in our proof of Theorem 5. This theorem is a simple extension of [69, Proposition 2] to the misspecified case, where estimators are searched over a function class that may not contain the true underlying function. For Theorem 5, we only need the partial result of Theorem 24 concerning the well-specified case. The more general result covering the misspecified case is only required for Theorem 7. However, here, we state the full result in order to avoid redundancy. We provide our proof of Theorem 24 in Appendix C.18.

Theorem 24. *Suppose that the data* $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$ *are generated according to the model*

$$y_i = f^*(\mathbf{x}^{(i)}) + \xi_i, \tag{C.23}$$

where $\mathbf{x}^{(i)}$ *are i.i.d. random variables with law* P *on* $\mathcal{X}$, ξ_i *are i.i.d. errors independent of* $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ *with mean zero and finite* L^q *norm for some* $q \geq 3$, *and* f^* *is an unknown real-valued regression function defined on* $\mathcal{X}$. *Let* $\mathcal{F}$ *be a countable collection of real-valued functions defined on* $\mathcal{X}$ *and* f_0 *be an element of* $\mathcal{F}$, *and assume that* $\|f - f_0\|_\infty \leq M$ *for every* $f \in \mathcal{F}$. *Also, let* $\hat{f}$ *be an estimator of* f^* *for which*

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 \right\}$$

with probability at least $1 - \epsilon$ *for some* $\epsilon \geq 0$. *Moreover, suppose that there exists* $t_n \geq 4\|f_0 - f^*\|_{P,2}$ *such that*

$$\mathbb{E}\Bigg[\sup_{\substack{f \in \mathcal{F} - \{f_0\} \\ \|f\|_{P,2} \leq r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i f(\mathbf{x}^{(i)}) \Big| \Bigg] \leq r\sqrt{n} t_n^2,$$

$$\mathbb{E}\Bigg[\sup_{\substack{f \in \mathcal{F} - \{f_0\} \\ \|f\|_{P,2} \leq r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f(\mathbf{x}^{(i)}) \Big| \Bigg] \leq r\sqrt{n} t_n^2,$$

and

$$\mathbb{E}\Bigg[\sup_{\substack{f \in \mathcal{F} - \{f_0\} \\ \|f\|_{P,2} \leq r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)}) \Big| \Bigg] \leq r\sqrt{n} t_n^2$$

for every $r \geq 1$. *Here,* $\|\cdot\|_{P,2}$ *is the norm defined by*

$$\|f\|_{P,2} = \left(\mathbb{E}_{\mathbf{X} \sim P}[f^2(\mathbf{X})]\right)^{1/2}$$

and ϵ_i are i.i.d. Rademacher variables independent of $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. Under these assumptions, there exists a universal positive constant C such that for every $t \geq t_n$,

$$\begin{aligned}\mathbb{P}(\|\hat{f} - f_0\|_{P,2} > t) \leq \epsilon + C \cdot \frac{1 + M^3}{t^3} \cdot t_n^3 + C \cdot \frac{\|\xi_1\|_2^3 + \|f_0 - f^*\|_\infty^3 + M^3}{n^{3/2} t^3} \\ + C \cdot \frac{M^3(\|\xi_1\|_q^3 n^{3/q} + \|f_0 - f^*\|_\infty^3 + M^3)}{n^3 t^6}.\end{aligned} \tag{C.24}$$

Remark 11. *The reason why $\mathcal{F}$ is assumed to be countable in Theorem 24 is to guarantee the measurability of the suprema. We can remove this condition if $\mathcal{F}$ is a subset of $C([0,1]^m)$.*

Suppose $\mathcal{F}$ is a subset of $C([0,1]^m)$ and $\mathcal{F}_0$ is a subset of $\mathcal{F} - \{f_0\}$. Because of our assumption that $f_0 \in \mathcal{F}$, $\mathcal{F}_0$ is a subset of $C([0,1]^m)$ as well. Also, since $(C([0,1]^m), \|\cdot\|_\infty)$ is a separable metric space, $\mathcal{F}_0$ is also separable with respect to the sup-norm. Let $\mathcal{G}_0$ be a countable dense subset of $\mathcal{F}_0$ with respect to the sup-norm. Then, we have

$$\sup_{f \in \mathcal{F}_0} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)}) \Big| = \sup_{g \in \mathcal{G}_0} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i g(\mathbf{x}^{(i)}) \Big|,$$
$$\sup_{f \in \mathcal{F}_0} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)}) \Big| = \sup_{g \in \mathcal{G}_0} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i g(\mathbf{x}^{(i)}) \Big|,$$

and

$$\sup_{f \in \mathcal{F}_0} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)}) \Big| = \sup_{g \in \mathcal{G}_0} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i g(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)}) \Big|$$

so that all the suprema are measurable. For these reasons, in this case, Theorem 24 is valid without the countability assumption.

Remark 12 (Well-specified Case)**.** *If f^* belongs to the function class $\mathcal{F}$, we can set $f_0 = f^*$. In this case, the conditions $t_n \geq 4\|f_0 - f^*\|_{P,2}$ and*

$$\mathbb{E}\Big[\sup_{\substack{f \in \mathcal{F} - \{f_0\} \\ \|f\|_{P,2} \leq r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)}) \Big| \Big] \leq r \sqrt{n} t_n^2$$

always hold, and (C.24) *is simplified as*

$$\mathbb{P}(\|\hat{f} - f^*\|_{P,2} > t) \leq \epsilon + C \cdot \frac{1 + M^3}{t^3} \cdot t_n^3 + C \cdot \frac{\|\xi_1\|_2^3 + M^3}{n^{3/2} t^3} + C \cdot \frac{M^3(\|\xi_1\|_q^3 n^{3/q} + M^3)}{n^3 t^6}.$$

This simplified version will be used for our proof of Theorem 5.

Recall that our estimator $\hat{f}_{n,V}^{d,s}$ is defined as a least squares estimator over

$$\big\{ f \in \mathcal{F}_{\infty-\mathrm{mars}}^{d,s} : V_{\mathrm{mars}}(f) \leq V \big\},$$

which is not uniformly bounded as required for $\mathcal{F}$ in Theorem 24. Theorem 24 is therefore not directly applicable to our estimator, but we can avoid this problem by using the fact that $\hat{f}^{d,s}_{n,V}$ is (uniformly) bounded in probability. Due to the boundedness of $\hat{f}^{d,s}_{n,V}$ as stated in Lemma 15 (proved in Appendix C.19), it is guaranteed that $\hat{f}^{d,s}_{n,V}$ is a least squares estimator over a uniformly bounded function class with high probability.

Lemma 15. *The estimator $\hat{f}^{d,s}_{n,V}$ satisfies*

$$\|\hat{f}^{d,s}_{n,V} - f^*\|_\infty = O_p(1).$$

We also use the following result from [69]. This theorem will help us bound

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le rt_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big]$$

in Theorem 24 using upper bounds of

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le rt_k}} \Big|\sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big]$$

for suitably chosen t_k, for $1 \le k \le n$.

Theorem 25 (Corollary 1 of [69]). *Let $\mathcal{F}_1, \ldots, \mathcal{F}_n$ be countable collections of real-valued functions defined on $\mathcal{X}$ for which $\mathcal{F}_k \supseteq \mathcal{F}_n$ for every $1 \le k \le n$. Suppose $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are permutation invariant random variables on $\mathcal{X}$ and $\xi_1, \ldots, \xi_n$ are i.i.d. mean-zero random variables independent of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$. Assume that there exist positive constants $r \ge 1$ and C such that*

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}_k} \Big|\sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] \le Ck^{1/r}$$

for every $1 \le k \le n$. Then, for every $q \ge 1$, we have

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}_n} \Big|\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] \le 4C\|\xi_1\|_{\min\{q,r\},1} \cdot n^{1/\min\{q,r\}},$$

where for each $p \ge 1$,

$$\|\xi_1\|_{p,1} := \int_0^\infty (\mathbb{P}(|\xi_1| > t))^{1/p}\, dt.$$

As in Theorem 24, the countability assumption in Theorem 25 can be dropped if $\mathcal{F}_k$ are subsets of $C([0,1]^m)$.

Proof of Theorem 5. Suppose $\epsilon > 0$. By Lemma 15, there exists $M > 0$ such that

$$\mathbb{P}\big(\|\hat{f}^{d,s}_{n,V} - f^*\|_\infty \leq M\big) \geq 1 - \epsilon$$

for sufficiently large n. For $V > 0$, let

$$\mathcal{F}(V, M) = \big\{f \in \mathcal{F}^{d,s}_{\infty-\text{mars}} : V_{\text{mars}}(f) \leq V \text{ and } \|f - f^*\|_\infty \leq M\big\}.$$

Then, it is clear from the definition of $\hat{f}^{d,s}_{n,V}$ that for sufficiently large n,

$$\hat{f}^{d,s}_{n,V} \in \underset{f \in \mathcal{F}(V,M)}{\operatorname{argmin}} \left\{\sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)})\big)^2\right\}$$

with probability at least $1 - \epsilon$. Also, we can see that Theorem 24 is valid for $\mathcal{F}(V, M)$ since $\mathcal{F}(V, M)$ is a subset of $C([0, 1]^d)$.

We first bound

$$\mathbb{E}\bigg[\sup_{\substack{f \in \mathcal{F}(V,M)-\{f^*\} \\ \|f\|_{p_0,2} \leq t}} \Big|\frac{1}{\sqrt{k}} \sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\bigg]$$

for $1 \leq k \leq n$ and $t > 0$. The following lemma (proved in Appendix C.20) helps us bound it using the bracketing entropy integral of $\mathcal{F}(V, M) - \{f^*\}$ (under the norm $\|\cdot\|_{p_0,2}$). Recall that the bracketing entropy integral of a function class $\mathcal{G}$ under a norm $\|\cdot\|$ is defined as

$$J_{[\,]}(t, \mathcal{G}, \|\cdot\|) = \int_0^t \sqrt{1 + \log N_{[\,]}(\epsilon, \mathcal{G}, \|\cdot\|)}\, d\epsilon \quad \text{for } t > 0.$$

Lemma 16. *There exists a universal positive constant C such that for every $t > 0$,*

$$\mathbb{E}\bigg[\sup_{\substack{f \in \mathcal{F}(V,M)-\{f^*\} \\ \|f\|_{p_0,2} \leq t}} \Big|\frac{1}{\sqrt{k}} \sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\bigg] \\ \leq C J_{[\,]}(t, \mathcal{F}(V, M) - \{f^*\}, \|\cdot\|_{p_0,2}) \cdot \left(1 + M \cdot \frac{J_{[\,]}(t, \mathcal{F}(V, M) - \{f^*\}, \|\cdot\|_{p_0,2})}{t^2\sqrt{k}}\right).$$

Also, the following lemma provides an upper bound on the bracketing entropy of $\mathcal{F}(V, M) - \{f^*\}$ (under the norm $\|\cdot\|_{p_0,2}$). We defer our proof of this lemma to Appendix C.21.

Lemma 17. *There exist positive constants C_s depending on s and $C_{B,d}$ depending on B and d such that*

$$\log N_{[\,]}(\epsilon, \mathcal{F}(V, M) - \{f^*\}, \|\cdot\|_{p_0,2}) \leq 2^d \log\left(2 + C_s \cdot \frac{M + V}{\epsilon}\right) \\ + C_{B,d}\left(2^{d+2} + \frac{2^{d+3}V}{\epsilon}\right)^{1/2}\left[\log\left(2^{d+2} + \frac{2^{d+3}V}{\epsilon}\right)\right]^{2(s-1)}$$

for every $V > 0, M > 0,$ and $\epsilon > 0$.

By Lemma 17 and the inequality $(x+y)^{1/2} \leq x^{1/2} + y^{1/2}$,

$$\begin{aligned}
J_{[\,]}(t, \mathcal{F}(V,M) - \{f^*\}, \|\cdot\|_{p_0,2}) &= \int_0^t \sqrt{1 + \log N_{[\,]}(\epsilon, \mathcal{F}(V,M) - \{f^*\}, \|\cdot\|_{p_0,2})}\, d\epsilon \\
&\leq t + 2^{d/2} \int_0^t \sqrt{\log\Big(2 + C_s \cdot \frac{M+V}{\epsilon}\Big)}\, d\epsilon \\
&\quad + C_{B,d} \int_0^t \Big(2^{d+2} + \frac{2^{d+3}V}{\epsilon}\Big)^{1/4} \Big[\log\Big(2^{d+2} + \frac{2^{d+3}V}{\epsilon}\Big)\Big]^{s-1} d\epsilon.
\end{aligned}$$

We use Lemma 13 again for bounding the above integrals. Lemma 13 implies that

$$\begin{aligned}
\int_0^t \sqrt{\log\Big(2 + C_s \cdot \frac{M+V}{\epsilon}\Big)}\, d\epsilon &\leq \int_0^t \sqrt{\log\Big(\frac{2t + C_s(M+V)}{\epsilon}\Big)}\, d\epsilon \\
&\leq Ct\Big[1 + 2\log\Big(2 + C_s \cdot \frac{M+V}{t}\Big)\Big] \leq C_s t + Ct \log\Big(2 + \frac{M+V}{t}\Big)
\end{aligned}$$

and that

$$\begin{aligned}
&\int_0^t \Big(2^{d+2} + \frac{2^{d+3}V}{\epsilon}\Big)^{1/4} \Big[\log\Big(2^{d+2} + \frac{2^{d+3}V}{\epsilon}\Big)\Big]^{s-1} d\epsilon \\
&\quad \leq \int_0^t \Big(\frac{2^{d+2}(t+2V)}{\epsilon}\Big)^{1/4} \Big[\log\Big(\frac{2^{d+2}(t+2V)}{\epsilon}\Big)\Big]^{s-1} d\epsilon \\
&\quad \leq C_s \big[2^{d+2}(t+2V)\big]^{1/4} t^{3/4} \Big[1 + \Big\{\log\Big(2^{d+2} + \frac{2^{d+3}V}{t}\Big)\Big\}^{s-1}\Big] \\
&\quad \leq C_d (t + V^{1/4} t^{3/4}) \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1} \\
&\quad \leq C_d t \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1} + C_d V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1}.
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
J_{[\,]}(t, \mathcal{F}(V,M) - \{f^*\}, \|\cdot\|_{p_0,2}) &\leq C_d t \log\Big(2 + \frac{M+V}{t}\Big) + C_{B,d} t \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1} \\
&\qquad + C_{B,d} V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1}.
\end{aligned}$$

Note that if $t \leq V$,

$$C_{B,d} t \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1} \leq C_{B,d} V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1}$$

and otherwise,

$$C_{B,d} t \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1} \leq C_{B,d} t \leq C_{B,d} t \log\Big(2 + \frac{M+V}{t}\Big).$$

Therefore, we have

$$J_{[\,]}(t, \mathcal{F}(V,M) - \{f^*\}, \|\cdot\|_{p_0,2}) \le C_{B,d} t \log\Big(2 + \frac{M+V}{t}\Big) + C_{B,d} V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1},$$

which leads to

$$\begin{aligned}\mathbb{E}\Big[\sup_{\substack{f \in \mathcal{F}(V,M)-\{f^*\} \\ \|f\|_{p_0,2} \le t}} \Big|\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] &\le C_{B,d} t \log\Big(2 + \frac{M+V}{t}\Big) \\ &+ C_{B,d} V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1} + C_{B,d} M k^{-1/2} \Big[\log\Big(2 + \frac{M+V}{t}\Big)\Big]^2 \\ &+ C_{B,d} M V^{1/4} t^{-1/4} k^{-1/2} \log\Big(2 + \frac{M+V}{t}\Big) \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1} \\ &+ C_{B,d} M V^{1/2} t^{-1/2} k^{-1/2} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{2(s-1)} =: \Psi_k(t)\end{aligned}$$

due to Lemma 16.

Note that $t \to \Psi_k(t)/t$ is monotonically decreasing. Also, observe that for

$$t \ge (1 + M^{1/2}) k^{-1/2},$$

we have

$$C_{B,d} t \log\Big(2 + \frac{M+V}{t}\Big) \le \frac{1}{5}\sqrt{k} t^2 \quad \text{if} \ \ t \ge C_{B,d} k^{-1/2} \log\Big(2 + \frac{(M+V)k^{1/2}}{1 + M^{1/2}}\Big),$$

$$C_{B,d} M k^{-1/2} \Big[\log\Big(2 + \frac{M+V}{t}\Big)\Big]^2 \le \frac{1}{5}\sqrt{k} t^2$$
$$\text{if} \ \ t \ge C_{B,d} M^{1/2} k^{-1/2} \log\Big(2 + \frac{(M+V)k^{1/2}}{1 + M^{1/2}}\Big),$$

$$C_{B,d} V^{1/4} t^{3/4} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1} \le \frac{1}{5}\sqrt{k} t^2 \quad \text{if} \ \ t \ge C_{B,d} V^{1/5} k^{-2/5} \Big[\log\Big(2 + \frac{V k^{1/2}}{1 + M^{1/2}}\Big)\Big]^{4(s-1)/5},$$

$$C_{B,d} M V^{1/4} t^{-1/4} k^{-1/2} \log\Big(2 + \frac{M+V}{t}\Big) \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{s-1} \le \frac{1}{5}\sqrt{k} t^2$$
$$\text{if} \ \ t \ge C_{B,d} M^{4/9} V^{1/9} k^{-4/9} \Big[\log\Big(2 + \frac{(M+V)k^{1/2}}{1 + M^{1/2}}\Big)\Big]^{4/9} \Big[\log\Big(2 + \frac{V k^{1/2}}{1 + M^{1/2}}\Big)\Big]^{4(s-1)/9},$$

and

$$C_{B,d} M V^{1/2} t^{-1/2} k^{-1/2} \Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{2(s-1)} \le \frac{1}{5}\sqrt{k} t^2$$
$$\text{if} \ \ t \ge C_{B,d} M^{2/5} V^{1/5} k^{-2/5} \Big[\log\Big(2 + \frac{V k^{1/2}}{1 + M^{1/2}}\Big)\Big]^{4(s-1)/5}.$$

Clearly, there exist some positive constants C_s and C such that

$$[\log(2+x)]^{4s/9} \le x^{4/45} \quad \text{for } x \ge C_s \quad \text{and} \quad \log(2+x) \le x^{1/5} \quad \text{for } x \ge C.$$

Thus, it follows that

$$\begin{aligned}
&C_{B,d}(1+M^{1/2})k^{-1/2}\log\Big(2+\frac{(M+V)k^{1/2}}{1+M^{1/2}}\Big)\\
&\quad\le C_{B,d}(1+M^{1/2})k^{-1/2}+C_{B,d}(1+M^{1/2})k^{-1/2}\Big(\frac{(M+V)k^{1/2}}{1+M^{1/2}}\Big)^{1/5}\\
&\quad\le C_{B,d}(1+M^{1/2})k^{-1/2}+C_{B,d}(1+M^{1/2})^{4/5}(M+V)^{1/5}k^{-2/5}\\
&\quad\le C_{B,d}(1+M^{1/2})^{4/5}\big[(1+M^{1/2})^{1/5}+(M+V)^{1/5}\big]k^{-2/5}
\end{aligned}$$

and

$$\begin{aligned}
&C_{B,d}M^{4/9}V^{1/9}k^{-4/9}\Big[\log\Big(2+\frac{(M+V)k^{1/2}}{1+M^{1/2}}\Big)\Big]^{4/9}\Big[\log\Big(2+\frac{Vk^{1/2}}{1+M^{1/2}}\Big)\Big]^{4(s-1)/9}\\
&\quad\le C_{B,d}M^{4/9}V^{1/9}k^{-4/9}\Big[\log\Big(2+\frac{(M+V)k^{1/2}}{1+M^{1/2}}\Big)\Big]^{4s/9}\\
&\quad\le C_{B,d}M^{4/9}V^{1/9}k^{-4/9}+C_{B,d}M^{4/9}V^{1/9}k^{-4/9}\Big(\frac{(M+V)k^{1/2}}{1+M^{1/2}}\Big)^{4/45}\\
&\quad\le C_{B,d}M^{4/9}V^{1/9}\Big[1+\Big(\frac{M+V}{1+M^{1/2}}\Big)^{4/45}\Big]k^{-2/5}.
\end{aligned}$$

For these reasons, for $k=1,\dots,n$, if we let

$$\begin{aligned}
t_k=\Big\{&C_{B,d}(1+M^{1/2})^{4/5}\big[(1+M^{1/2})^{1/5}+(M+V)^{1/5}\big]\\
&+C_{B,d}M^{4/9}V^{1/9}\Big[1+\Big(\frac{M+V}{1+M^{1/2}}\Big)^{4/45}\Big]\\
&+C_{B,d}(1+M^{1/2})^{4/5}V^{1/5}\Big[\log\Big(2+\frac{Vn^{1/2}}{1+M^{1/2}}\Big)\Big]^{4(s-1)/5}\Big\}\cdot k^{-2/5}
\end{aligned}$$

then

$$\Psi_k(t_k)\le\sqrt{k}t_k^2.$$

Hence, for $r\ge 1$, we have

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le rt_k}}\Big|\sum_{i=1}^k\epsilon_i f(\mathbf{x}^{(i)})\Big|\Big]\le\sqrt{k}\cdot\Psi_k(rt_k)\le r\sqrt{k}\cdot\Psi_k(t_k)\le rkt_k^2,$$

where the second inequality follows from that $t\to\Psi_k(t)/t$ is monotonically decreasing.

Recall that we assume $\|\xi_1\|_{5,1} < \infty$. Using Theorem 25, we can thus show that

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le rt_n}} \Big|\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] \le 4\|\xi_1\|_{5,1}\cdot rnt_n^2$$

for $r \ge 1$. Therefore, if we redefine t_n as $(1+4\|\xi_1\|_{5,1})t_n$, then

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le rt_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] \le r\sqrt{n}t_n^2$$

and

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le rt_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] \le r\sqrt{n}t_n^2.$$

As a result, by Theorem 24 (and Remark 12),

$$\begin{aligned}\mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f^*\|_{p_0,2} > t\big) &\le \epsilon + C\cdot\frac{1+M^3}{t^3}\cdot t_n^3 \\ &\qquad + C\cdot\frac{\|\xi_1\|_2^3+M^3}{n^{3/2}t^3} + C\cdot\frac{M^3(\|\xi_1\|_5^3 n^{3/5}+M^3)}{n^3t^6}\end{aligned}$$

provided that n is sufficiently large and $t \ge t_n$.

Observe that if $s = 1$,

$$\begin{aligned}t_n &\le (1+4\|\xi_1\|_{5,1})\cdot\Big\{C_{B,d}(1+M^{1/2})^{4/5}\big[(1+M^{1/2})^{1/5}+(M+V)^{1/5}\big] \\ &\qquad\qquad + C_{B,d}M^{4/9}V^{1/9}\Big[1+\Big(\frac{M+V}{1+M^{1/2}}\Big)^{4/45}\Big]\Big\}\cdot n^{-2/5} \\ &\le (1+4\|\xi_1\|_{5,1})\cdot C_{B,d,V,M}\cdot n^{-2/5},\end{aligned}$$

and otherwise ($s \ge 2$),

$$\begin{aligned}t_n &= (1+4\|\xi_1\|_{5,1})\cdot C_{B,d}(1+M^{1/2})^{4/5}V^{1/5}\Big[\log\Big(2+\frac{Vn^{1/2}}{1+M^{1/2}}\Big)\Big]^{4(s-1)/5}\cdot n^{-2/5} + O(n^{-2/5}) \\ &\le (1+4\|\xi_1\|_{5,1})\cdot C_{B,d,V,M}\Big[\log\Big(2+\frac{Vn^{1/2}}{1+M^{1/2}}\Big)\Big]^{4(s-1)/5}\cdot n^{-2/5} + O(n^{-2/5}),\end{aligned}$$

where the multiplicative constant underlying $O(\cdot)$ depends on B, d, V, M, and the moments of ξ_i. Thus, if

$$K \ge 2(1+4\|\xi_1\|_{5,1})\cdot C_{B,d,V,M},$$

then

$$Kn^{-2/5}(\log n)^{4(s-1)/5} \ge t_n$$

for sufficiently large n. Hence, for such K, we have

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\big(\|\hat{f}^{d,s}_{n,V} - f^*\|_{p_0,2} > Kn^{-2/5}(\log n)^{4(s-1)/5}\big) \\
&\quad\le \epsilon + \limsup_{n\to\infty} \Bigg[C \cdot \frac{(1+M^3)t_n^3}{K^3 n^{-6/5}(\log n)^{12(s-1)/5}} + C \cdot \frac{\|\xi_1\|_2^3 + M^3}{K^3 n^{3/10}(\log n)^{12(s-1)/5}} \\
&\qquad\qquad + C \cdot \frac{M^3(\|\xi_1\|_5^3 n^{3/5} + M^3)}{K^6 n^{3/5}(\log n)^{24(s-1)/5}}\Bigg] \\
&\quad\le \epsilon + (1 + 4\|\xi_1\|_{5,1})^3 \cdot C_{B,d,V,M} \cdot \frac{1}{K^3} + \|\xi_1\|_5^3 \cdot CM^3 \cdot \frac{1}{K^6}.
\end{aligned}$$

Consequently, we can find $K > 0$ for which

$$\mathbb{P}\big(\|\hat{f}^{d,s}_{n,V} - f^*\|_{p_0,2} > Kn^{-2/5}(\log n)^{4(s-1)/5}\big) < 2\epsilon$$

for sufficiently large n. $\square$

C.11 Proof of Theorem 6

We prove Theorem 6 by using [55, Theorem 1.1] and ideas in the proof of Theorem 1.2 of the same paper. For a positive integer m and $R > 0$, let $\mathcal{H}^+_m(R)$ be the collection of all functions on $[0,1]^m$ of the form

$$(x_1, \dots, x_m) \mapsto \int \mathbf{1}(x_1 \ge t_1) \cdots \mathbf{1}(x_m \ge t_m)\, d\nu(\mathbf{t}) = \nu\Big(\prod_{j=1}^m [0, x_j]\Big)$$

where ν is a (Borel) measure on $[0,1]^m$ with $\nu([0,1]^m) \le R$. [55, Theorem 1.1] (restated below) provides an upper bound on the bracketing entropy of $\mathcal{H}^+_m(R)$ under the L^p norm. Also, ideas in the proof of [55, Theorem 1.2] help us obtain an upper bound on the bracketing entropy of $\mathcal{D}_m$ under the L^2 norm from those of $\mathcal{H}^+_m(R)$ under the L^1 and L^2 norms.

Theorem 26 (Theorem 1.1 of [55])**.** *Suppose that $p \ge 1$ and m is a positive integer. Then, there exists a positive constant $C_{m,p}$ depending on m and p such that*

$$\log N_{[\,]}(\epsilon, \mathcal{H}^+_m(R), \|\cdot\|_p) \le C_{m,p} \cdot \frac{R}{\epsilon}\Big|\log \frac{R}{\epsilon}\Big|^{2(m-1)} \quad \textit{for every } \epsilon > 0.$$

Proof of Theorem 6. Assume that $0 < \epsilon < 2$. If $\epsilon \ge 2$, then

$$\log N_{[\,]}(\epsilon, \mathcal{D}_m, \|\cdot\|_2) = 0$$

because

$$-1 \le \int (x_1 - t_1)_+ \cdots (x_m - t_m)_+ \, d\nu(\mathbf{t}) \le 1$$

for every signed measure ν on $[0,1]^m$ with total variation $|\nu|([0,1]^m) \leq 1$. Denote by $\mathcal{D}_m^+$ the subcollection of $\mathcal{D}_m$ consisting of all functions on $[0,1]^m$ of the form

$$\int (x_1 - t_1)_+ \cdots (x_m - t_m)_+ \, d\nu(\mathbf{t})$$

where ν is a measure on $[0,1]^m$ with $\nu([0,1]^m) \leq 1$. Then, since

$$\mathcal{D}_m \subseteq \mathcal{D}_m^+ - \mathcal{D}_m^+,$$

it follows that

$$\log N_{[\,]}(\epsilon, \mathcal{D}_m, \|\cdot\|_2) \leq 2 \log N_{[\,]}\Big(\frac{\epsilon}{2}, \mathcal{D}_m^+, \|\cdot\|_2\Big). \tag{C.25}$$

Recall that $\mathcal{T}_m : L^2([0,1]^m) \to C([0,1]^m)$ is the m-dimensional integral operator defined by

$$\mathcal{T}_m f(\mathbf{x}) = \int_{\prod_{j=1}^m [0,x_j]} f(\mathbf{t}) \, d\mathbf{t} \quad \text{for } \mathbf{x} \in [0,1]^m.$$

Since

$$\int_{[0,1]^m} (x_1 - t_1)_+ \cdots (x_m - t_m)_+ \, d\nu(\mathbf{t}) = \int_{\prod_{j=1}^m [0,x_j]} \nu\bigg(\prod_{j=1}^m [0, s_j]\bigg) \, d\mathbf{s} \quad \text{for } \mathbf{x} \in [0,1]^m,$$

for every measure ν on $[0,1]^m$ with $\nu([0,1]^m) \leq 1$, we have

$$\mathcal{D}_m^+ = \mathcal{T}_m(\mathcal{H}_m^+) \quad \text{where } \mathcal{H}_m^+ := \mathcal{H}_m^+(1).$$

Now, fix $\delta > 0$, which will be specified later, and let $K = N(\delta, \mathcal{H}_m^+, \|\cdot\|_1)$. Note that by Theorem 26,

$$\log K = \log N(\delta, \mathcal{H}_m^+, \|\cdot\|_1) \leq \log N_{[\,]}(\delta, \mathcal{H}_m^+, \|\cdot\|_1) \leq C_m \cdot \frac{1}{\delta} \bigg| \log \frac{1}{\delta} \bigg|^{2(m-1)}. \tag{C.26}$$

Also, let $h_1, \dots, h_K$ be a δ-net of $\mathcal{H}_m^+$ with respect to the L^1 norm, i.e.,

$$\mathcal{H}_m^+ \subseteq \bigcup_{k=1}^K B_{\mathcal{H}_m^+}(\delta, h_k, \|\cdot\|_1)$$

where $B_{\mathcal{H}_m^+}(\delta, h_k, \|\cdot\|_1) := \{h \in \mathcal{H}_m^+ : \|h - h_k\|_1 \leq \delta\}$. Next, for $k = 1, \dots, K$, let

$$\mathcal{I}_k^+ = \big\{\mathcal{T}_m((h - h_k)_+) : h \in \mathcal{H}_m^+ \text{ and } \|h - h_k\|_1 \leq \delta\big\}$$

and

$$\mathcal{I}_k^- = \big\{\mathcal{T}_m((h_k - h)_+) : h \in \mathcal{H}_m^+ \text{ and } \|h - h_k\|_1 \leq \delta\big\}.$$

Observe that $\mathcal{I}_k^+, \mathcal{I}_k^- \subseteq \mathcal{H}_m^+(\delta)$. Indeed, if $h \in \mathcal{H}_m^+$ and $\|h - h_k\|_1 \leq \delta$, then for a measure ν on $[0,1]^m$ defined by

$$d\nu(\mathbf{t}) = (h(\mathbf{t}) - h_k(\mathbf{t}))_+ \, d\mathbf{t},$$

we have

$$\nu([0,1]^m) = \int_{[0,1]^m} (h(\mathbf{t}) - h_k(\mathbf{t}))_+ \, d\mathbf{t} \leq \|h - h_k\|_1 \leq \delta$$

and

$$\mathcal{T}_m((h - h_k)_+)(\mathbf{x}) = \int_{\prod_{j=1}^m [0,x_j]} (h(\mathbf{t}) - h_k(\mathbf{t}))_+ \, d\mathbf{t} = \int_{\prod_{j=1}^m [0,x_j]} d\nu(\mathbf{t}) = \nu\Big(\prod_{j=1}^m [0, x_j]\Big)$$

for all $\mathbf{x} \in [0,1]^m$. Also, because $\mathcal{D}_m^+ = \mathcal{T}_m(\mathcal{H}_m^+)$, $h_1, \ldots, h_K$ form a δ-net of $\mathcal{H}_m^+$ with respect to the L^1 norm, and

$$\mathcal{T}_m(h) = \mathcal{T}_m\big((h - h_k)_+ - (h_k - h)_+ + h_k\big) = \mathcal{T}_m((h - h_k)_+) - \mathcal{T}_m((h_k - h)_+) + \mathcal{T}_m(h_k)$$

for $h \in \mathcal{H}_m^+$ and $k = 1, \ldots, K$, it follows that

$$\mathcal{D}_m^+ \subseteq \bigcup_{k=1}^K \big(\mathcal{I}_k^+ - \mathcal{I}_k^- + \{\mathcal{T}_m(h_k)\}\big).$$

Thus,

$$N_{[\,]}(\epsilon, \mathcal{D}_m^+, \|\cdot\|_2) \leq \sum_{k=1}^K N_{[\,]}\Big(\frac{\epsilon}{2}, \mathcal{I}_k^+, \|\cdot\|_2\Big) \cdot N_{[\,]}\Big(\frac{\epsilon}{2}, \mathcal{I}_k^-, \|\cdot\|_2\Big),$$

which together with (C.26) and Theorem 26 implies that

$$\begin{aligned} \log N_{[\,]}(\epsilon, \mathcal{D}_m^+, \|\cdot\|_2) &\leq \log K + 2 \log N_{[\,]}\Big(\frac{\epsilon}{2}, \mathcal{H}_m^+(\delta), \|\cdot\|_2\Big) \\ &\leq C_m \cdot \frac{1}{\delta}\Big|\log \frac{1}{\delta}\Big|^{2(m-1)} + C_m \cdot \frac{2\delta}{\epsilon}\Big|\log \frac{2\delta}{\epsilon}\Big|^{2(m-1)}. \end{aligned}$$

Therefore, with the choice of $\delta = (\epsilon/2)^{1/2}$, we can derive that

$$\log N_{[\,]}(\epsilon, \mathcal{D}_m^+, \|\cdot\|_2) \leq C_m \cdot \Big(\frac{2}{\epsilon}\Big)^{1/2}\Big|\log \frac{2}{\epsilon}\Big|^{2(m-1)}.$$

By (C.25), we can conclude that

$$\log N_{[\,]}(\epsilon, \mathcal{D}_m, \|\cdot\|_2) \leq C_m \cdot \Big(\frac{4}{\epsilon}\Big)^{1/2}\Big|\log \frac{4}{\epsilon}\Big|^{2(m-1)}.$$

□

C.12 Proof of Theorem 7

Proof of Theorem 7. Suppose we are given $\epsilon > 0$. By repeating our argument in Lemma 15, we can show that there exists $M > 0$ independent of N such that

$$\mathbb{P}\big(\|\widetilde{f}^{d,s}_{n,V} - f^*\|_\infty \le M\big) \ge 1 - \epsilon$$

for sufficiently large n. Recall that we let $\mathcal{F}_{\text{approx}}(V)$ be the collection of all functions $f_{c,\{\nu_S\}} \in \mathcal{F}^{d,s}_{\infty-\text{mars}}$ with $V_{\text{mars}}(f_{c,\{\nu_S\}}) \le V$ where ν_S is supported on $(\prod_{j\in S} \widetilde{\mathcal{V}}_j) \cap [0,1)^{|S|}$ for each $S \subseteq [d]$ with $0 < |S| \le s$. Due to Lemma 14, there exists $f_0 \in \mathcal{F}_{\text{approx}}(V)$ such that

$$\|f_0 - f^*\|_\infty \le \frac{2V}{N} \quad \text{and} \quad \|f_0 - f^*\|_{p_0,2} \le C_d\Big(\frac{B}{N^3}\Big)^{1/2} V.$$

Thus, by the triangle inequality, we have

$$\mathbb{P}\Big(\|\widetilde{f}^{d,s}_{n,V} - f_0\|_\infty \le M + \frac{2V}{N}\Big) \ge 1 - \epsilon$$

for sufficiently large n. For $V > 0$, let

$$\mathcal{F}\Big(V, M + \frac{2V}{N}\Big) = \Big\{f \in \mathcal{F}^{d,s}_{\infty-\text{mars}} : V_{\text{mars}}(f) \le V \text{ and } \|f - f_0\|_\infty \le M + \frac{2V}{N}\Big\}$$

as in the proof of Theorem 5, and let

$$\mathcal{F}_{\text{approx}}\Big(V, M + \frac{2V}{N}\Big) = \Big\{f \in \mathcal{F}_{\text{approx}}(V) : \|f - f_0\|_\infty \le M + \frac{2V}{N}\Big\}.$$

Then, from the definition of $\widetilde{f}^{d,s}_{n,V}$, we can see that

$$\widetilde{f}^{d,s}_{n,V} \in \underset{f \in \mathcal{F}_{\text{approx}}(V, M+2V/N)}{\operatorname{argmin}} \Big\{\sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)})\big)^2\Big\}$$

with probability at least $1 - \epsilon$.

Through the same argument as in the proof of Theorem 5 with f^* replaced by f_0 and M with $\widetilde{M} := M + 2V/N$, we can show that for

$$\begin{aligned} t_n = (1 + 4\|\xi_1\|_{5,1})\Big\{ & C_{B,d}(1 + \widetilde{M}^{1/2})^{4/5}\big[(1 + \widetilde{M}^{1/2})^{1/5} + (\widetilde{M} + V)^{1/5}\big] \\ & + C_{B,d}\widetilde{M}^{4/9}V^{1/9}\Big[1 + \Big(\frac{\widetilde{M} + V}{1 + \widetilde{M}^{1/2}}\Big)^{4/45}\Big] \\ & + C_{B,d}(1 + \widetilde{M}^{1/2})^{4/5}V^{1/5}\Big[\log\Big(2 + \frac{Vn^{1/2}}{1 + \widetilde{M}^{1/2}}\Big)\Big]^{4(s-1)/5}\Big\} \cdot n^{-2/5}, \end{aligned}$$

we have

$$\mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}_{\text{approx}}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le rt_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\Bigg] \le \mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le rt_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\Bigg] \le r\sqrt{n}t_n^2$$

and

$$\mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}_{\text{approx}}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le rt_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Bigg] \le \mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le rt_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Bigg] \le r\sqrt{n}t_n^2.$$

Hence, to apply Theorem 24, it suffices to find an upper bound on

$$\mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}_{\text{approx}}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le t}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\cdot(f_0-f^*)(\mathbf{x}^{(i)})\Big|\Bigg] \quad \text{for } t>0.$$

First, by making minor modifications to the proof of Lemma 16, we can prove that

$$\begin{aligned}
&\mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}_{\text{approx}}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le t}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\cdot(f_0-f^*)(\mathbf{x}^{(i)})\Big|\Bigg]\\
&\le \mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le t}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\cdot(f_0-f^*)(\mathbf{x}^{(i)})\Big|\Bigg]\\
&\le C\|f_0-f^*\|_\infty\cdot J_{[\,]}\big(t,\mathcal{F}(V,\widetilde{M})-\{f_0\},\|\cdot\|_{p_0,2}\big)\\
&\qquad\qquad\qquad\cdot\left(1+\widetilde{M}\cdot\frac{J_{[\,]}\big(t,\mathcal{F}(V,\widetilde{M})-\{f_0\},\|\cdot\|_{p_0,2}\big)}{t^2\sqrt{n}}\right)\\
&\le C\cdot\frac{V}{N}\cdot J_{[\,]}\big(t,\mathcal{F}(V,\widetilde{M})-\{f_0\},\|\cdot\|_{p_0,2}\big)\cdot\left(1+\widetilde{M}\cdot\frac{J_{[\,]}\big(t,\mathcal{F}(V,\widetilde{M})-\{f_0\},\|\cdot\|_{p_0,2}\big)}{t^2\sqrt{n}}\right).
\end{aligned}$$

Next, repeating the same computation as in the proof of Theorem 5, we can derive that

$$\begin{aligned}
&\mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}_{\text{approx}}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le t}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\cdot(f_0-f^*)(\mathbf{x}^{(i)})\Big|\Bigg]\\
&\le\frac{V}{N}\cdot\Bigg\{C_{B,d}t\log\left(2+\frac{\widetilde{M}+V}{t}\right)+C_{B,d}V^{1/4}t^{3/4}\left[\log\left(2+\frac{V}{t}\right)\right]^{s-1}\\
&\qquad+C_{B,d}\widetilde{M}n^{-1/2}\left[\log\left(2+\frac{\widetilde{M}+V}{t}\right)\right]^2
\end{aligned}$$

$$+ C_{B,d}\widetilde{M}V^{1/4}t^{-1/4}n^{-1/2}\log\Big(2+\frac{\widetilde{M}+V}{t}\Big)\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{s-1}$$
$$\left. + C_{B,d}\widetilde{M}V^{1/2}t^{-1/2}n^{-1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{2(s-1)}\right\} =: \widetilde{\Psi}_n(t)$$

for $t > 0$. Note that for

$$t \geq \Big(\frac{V}{N}\Big)^{1/2}\Big(\frac{V}{N}+\widetilde{M}\Big)^{1/2}n^{-1/2},$$

we have

$$C_{B,d}\cdot\frac{V}{N}\cdot t\log\Big(2+\frac{\widetilde{M}+V}{t}\Big) \leq \frac{1}{5}\sqrt{n}t^2 \quad \text{if } \ t \geq C_{B,d}\cdot\frac{V}{N}\cdot n^{-1/2}\log\bigg(2+\frac{N(\widetilde{M}+V)n^{1/2}}{V^{1/2}(N\widetilde{M}+V)^{1/2}}\bigg),$$

$$C_{B,d}\cdot\frac{V}{N}\cdot\widetilde{M}n^{-1/2}\Big[\log\Big(2+\frac{\widetilde{M}+V}{t}\Big)\Big]^2 \leq \frac{1}{5}\sqrt{n}t^2$$
$$\text{if } \ t \geq C_{B,d}\Big(\frac{V}{N}\Big)^{1/2}\widetilde{M}^{1/2}n^{-1/2}\log\bigg(2+\frac{N(\widetilde{M}+V)n^{1/2}}{V^{1/2}(N\widetilde{M}+V)^{1/2}}\bigg),$$

$$C_{B,d}\cdot\frac{V}{N}\cdot V^{1/4}t^{3/4}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{s-1} \leq \frac{1}{5}\sqrt{n}t^2$$
$$\text{if } \ t \geq C_{B,d}\Big(\frac{V}{N}\Big)^{4/5}V^{1/5}n^{-2/5}\bigg[\log\bigg(2+\frac{NV^{1/2}n^{1/2}}{(N\widetilde{M}+V)^{1/2}}\bigg)\bigg]^{4(s-1)/5},$$

$$C_{B,d}\cdot\frac{V}{N}\cdot\widetilde{M}V^{1/4}t^{-1/4}n^{-1/2}\log\Big(2+\frac{\widetilde{M}+V}{t}\Big)\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{s-1} \leq \frac{1}{5}\sqrt{n}t^2$$
$$\text{if } \ t \geq C_{B,d}\Big(\frac{V}{N}\Big)^{4/9}\widetilde{M}^{4/9}V^{1/9}n^{-4/9}\bigg[\log\bigg(2+\frac{N(\widetilde{M}+V)n^{1/2}}{V^{1/2}(N\widetilde{M}+V)^{1/2}}\bigg)\bigg]^{4/9}$$
$$\cdot\bigg[\log\bigg(2+\frac{NV^{1/2}n^{1/2}}{(N\widetilde{M}+V)^{1/2}}\bigg)\bigg]^{4(s-1)/9},$$

and

$$C_{B,d}\cdot\frac{V}{N}\cdot\widetilde{M}V^{1/2}t^{-1/2}n^{-1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{2(s-1)} \leq \frac{1}{5}\sqrt{n}t^2$$
$$\text{if } \ t \geq C_{B,d}\Big(\frac{V}{N}\Big)^{2/5}\widetilde{M}^{2/5}V^{1/5}n^{-2/5}\bigg[\log\bigg(2+\frac{NV^{1/2}n^{1/2}}{(N\widetilde{M}+V)^{1/2}}\bigg)\bigg]^{4(s-1)/5}.$$

Thus, if we let

$$\widetilde{t}_n = \max\bigg\{C_{B,d}\Big(\frac{V}{N}\Big)^{1/2}\Big(\frac{V}{N}+\widetilde{M}\Big)^{1/2}n^{-1/2}\log\bigg(2+\frac{N(\widetilde{M}+V)n^{1/2}}{V^{1/2}(N\widetilde{M}+V)^{1/2}}\bigg),$$

$$C_{B,d}\Big(\frac{V}{N}\Big)^{4/9}\widetilde{M}^{4/9}V^{1/9}n^{-4/9}\bigg[\log\bigg(2+\frac{N(\widetilde{M}+V)n^{1/2}}{V^{1/2}(N\widetilde{M}+V)^{1/2}}\bigg)\bigg]^{4/9} \cdot\bigg[\log\bigg(2+\frac{NV^{1/2}n^{1/2}}{(N\widetilde{M}+V)^{1/2}}\bigg)\bigg]^{4(s-1)/9},$$

$$C_{B,d}\Big(\frac{V}{N}\Big)^{2/5}\Big(\frac{V}{N}+\widetilde{M}\Big)^{2/5}V^{1/5}n^{-2/5}\bigg[\log\bigg(2+\frac{NV^{1/2}n^{1/2}}{(N\widetilde{M}+V)^{1/2}}\bigg)\bigg]^{4(s-1)/5}\bigg\},$$

then

$$\widetilde{\Psi}_n(\widetilde{t}_n) \le \sqrt{n}\widetilde{t}_n^2.$$

This implies that

$$\mathbb{E}\bigg[\sup_{\substack{f\in\mathcal{F}_{\text{approx}}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le r\widetilde{t}_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\cdot(f_0-f^*)(\mathbf{x}^{(i)})\Big|\bigg] \le \widetilde{\Psi}_n(r\widetilde{t}_n) \le r\widetilde{\Psi}_n(\widetilde{t}_n) \le r\sqrt{n}\widetilde{t}_n^2$$

for every $r \ge 1$. Here, the second inequality follows from the fact that $t \to \widetilde{\Psi}_n(t)/t$ is monotonically decreasing.

Now, let

$$\bar{t}_n = 4\|f_0 - f^*\|_{p_0,2} + t_n + \widetilde{t}_n.$$

Then, it follows that

$$\mathbb{E}\bigg[\sup_{\substack{f\in\mathcal{F}_{\text{approx}}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le r\bar{t}_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\bigg] \le \Big(\frac{r\bar{t}_n}{t_n}\Big)\sqrt{n}t_n^2 = r\sqrt{n}t_n\bar{t}_n \le r\sqrt{n}\bar{t}_n^2$$

for every $r \ge 1$. Similarly, we can show that for $r \ge 1$,

$$\mathbb{E}\bigg[\sup_{\substack{f\in\mathcal{F}_{\text{approx}}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le r\bar{t}_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\bigg] \le r\sqrt{n}\bar{t}_n^2$$

and

$$\mathbb{E}\bigg[\sup_{\substack{f\in\mathcal{F}_{\text{approx}}(V,\widetilde{M})-\{f_0\}\\ \|f\|_{p_0,2}\le r\bar{t}_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\cdot(f_0-f^*)(\mathbf{x}^{(i)})\Big|\bigg] \le r\sqrt{n}\bar{t}_n^2.$$

Therefore, by Theorem 24 and the fact that

$$\|f_0 - f^*\|_\infty \le \frac{2V}{N},$$

we have

$$\mathbb{P}\big(\|\widetilde{f}_{n,V}^{d,s} - f_0\|_{p_0,2} > t\big) \le \epsilon + C\cdot\frac{1+\widetilde{M}^3}{t^3}\cdot\bar{t}_n^3 + C\cdot\frac{\|\xi_1\|_2^3 + (2V/N)^3 + \widetilde{M}^3}{n^{3/2}t^3}$$

$$+ C \cdot \frac{\widetilde{M}^3(\|\xi_1\|_5^3 n^{3/5} + (2V/N)^3 + \widetilde{M}^3)}{n^3 t^6}$$

provided that n is sufficiently large and $t \geq \bar{t}_n$. Using

$$\|f_0 - f^*\|_{p_0,2} \leq C_d\Big(\frac{B}{N^3}\Big)^{1/2} V,$$

we can derive from this result that

$$\begin{aligned}
\mathbb{P}\big(\|\widetilde{f}^{d,s}_{n,V} - f^*\|_{p_0,2} > t\big) &\leq \mathbb{P}\Big(\|\widetilde{f}^{d,s}_{n,V} - f_0\|_{p_0,2} > t - C_d\Big(\frac{B}{N^3}\Big)^{1/2} V\Big) \\
&\leq \epsilon + C \cdot \frac{1 + (M + 2V/N)^3}{(t - C_d(B/N^3)^{1/2}V)^3} \cdot \bar{t}_n^3 + C \cdot \frac{\|\xi_1\|_2^3 + (M + 2V/N)^3}{n^{3/2}(t - C_d(B/N^3)^{1/2}V)^3} \\
&\quad + C \cdot \frac{(M + 2V/N)^3(\|\xi_1\|_5^3 n^{3/5} + (M + 2V/N)^3)}{n^3(t - C_d(B/N^3)^{1/2}V)^6}
\end{aligned}$$

provided that n is sufficiently large and

$$t \geq C_d\Big(\frac{B}{N^3}\Big)^{1/2} V + \bar{t}_n.$$

Because $N = \Omega(n^{4/15})$,

$$C_d\Big(\frac{B}{N^3}\Big)^{1/2} V + \bar{t}_n \leq (1 + 4\|\xi_1\|_{5,1}) \cdot C_{B,d,V,M} n^{-2/5} + o(n^{-2/5}),$$

if $s = 1$, and otherwise ($s \geq 2$),

$$\begin{aligned}
C_d\Big(\frac{B}{N^3}\Big)^{1/2} V + \bar{t}_n &\leq (1 + 4\|\xi_1\|_{5,1}) \cdot C_{B,d}(1 + M^{1/2})^{4/5} V^{1/5} \\
&\qquad \cdot \Big[\log\Big(2 + \frac{Vn^{1/2}}{1 + M^{1/2}}\Big)\Big]^{4(s-1)/5} \cdot n^{-2/5} + O(n^{-2/5}) \\
&\leq (1 + 4\|\xi_1\|_{5,1}) \cdot C_{B,d,V,M}\Big[\log\Big(2 + \frac{Vn^{1/2}}{1 + M^{1/2}}\Big)\Big]^{4(s-1)/5} \cdot n^{-2/5} + O(n^{-2/5}),
\end{aligned}$$

where the constants underlying $o(\cdot)$ and $O(\cdot)$ depend on B, d, V, M, and the moments of ξ_i. Hence, if

$$K \geq 2(1 + 4\|\xi_1\|_{5,1}) \cdot C_{B,d,V,M},$$

then

$$Kn^{-2/5}(\log n)^{4(s-1)/5} \geq C_d\Big(\frac{B}{N^3}\Big)^{1/2} V + \bar{t}_n$$

for sufficiently large n. For such K, again because $N = \Omega(n^{4/15})$, we have that

$$\limsup_{n\to\infty} \mathbb{P}\big(\|\widetilde{f}^{d,s}_{n,V} - f^*\|_{p_0,2} > Kn^{-2/5}(\log n)^{4(s-1)/5}\big)$$

$$\leq \epsilon + \limsup_{n\to\infty} \Big[C \cdot \frac{1+(M+2V/N)^3}{(Kn^{-2/5}(\log n)^{4(s-1)/5} - C_d(B/N^3)^{1/2}V)^3} \cdot \bar{t}_n^3 + C \cdot \frac{\|\xi_1\|_2^3 + (M+2V/N)^3}{n^{3/2}(Kn^{-2/5}(\log n)^{4(s-1)/5} - C_d(B/N^3)^{1/2}V)^3} + C \cdot \frac{(M+2V/N)^3(\|\xi_1\|_5^3 n^{3/5} + (M+2V/N)^3)}{n^3(Kn^{-2/5}(\log n)^{4(s-1)/5} - C_d(B/N^3)^{1/2}V)^6} \Big]$$

$$\leq \epsilon + (1+4\|\xi_1\|_{5,1})^3 \cdot C_{B,d,V,M} \cdot \frac{1}{K^3} + \|\xi_1\|_5^3 \cdot CM^3 \cdot \frac{1}{K^6}.$$

As a result, we can find $K > 0$ that satisfies

$$\mathbb{P}\big(\|\widetilde{f}^{d,s}_{n,V} - f^*\|_{p_0,2} > Kn^{-2/5}(\log n)^{4(s-1)/5}\big) < 2\epsilon$$

for sufficiently large n. □

C.13 Proof of Theorem 8

Here, we almost follow the proof of [46, Theorem 4.6], which itself obtained the ideas from [13, Section 4]. As in [46], we exploit Assouad's lemma in the following form.

Lemma 18 (Lemma 24.3 of [149]). *Let q be a positive integer. Suppose for each $\boldsymbol{\eta} \in \{-1,1\}^q$, we have $f_{\boldsymbol{\eta}} \in \mathcal{F}^{d,s}_{\infty-mars}$ with $V_{mars}(f_{\boldsymbol{\eta}}) \leq V$. Then, the minimax risk $\mathfrak{M}^{d,s}_{n,V}$ satisfies*

$$\mathfrak{M}^{d,s}_{n,V} \geq \frac{q}{8} \cdot \min_{\boldsymbol{\eta}\neq\boldsymbol{\eta}'} \frac{\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|^2_{p_0,2}}{H(\boldsymbol{\eta},\boldsymbol{\eta}')} \cdot \min_{H(\boldsymbol{\eta},\boldsymbol{\eta}')=1} \left(1 - \sqrt{\frac{1}{2}\mathbb{E}\big[K(\mathbb{P}_{f_{\boldsymbol{\eta}}}, \mathbb{P}_{f_{\boldsymbol{\eta}'}})\big]}\right),$$

where $H(\cdot,\cdot)$ denotes the Hamming distance $H(\boldsymbol{\eta},\boldsymbol{\eta}') := \sum_{j=1}^q \mathbf{1}(\eta_j \neq \eta'_j)$, $\mathbb{P}_f$ is the probability distribution of $(y_1,\ldots,y_n)$ given $(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)})$ in (3.25) when $f^* = f$, and $K(\cdot,\cdot)$ indicates the Kullback divergence between two probability distributions.

Remark 13. *What [149, Lemma 24.3] actually tells us in the setting of Lemma 18 is that*

$$\max_{\boldsymbol{\eta}} \mathbb{E}_{f_{\boldsymbol{\eta}}}\|\hat{f}_n - f_{\boldsymbol{\eta}}\|^2_{p_0,2} \geq \frac{q}{8} \cdot \min_{\boldsymbol{\eta}\neq\boldsymbol{\eta}'} \frac{\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|^2_{p_0,2}}{H(\boldsymbol{\eta},\boldsymbol{\eta}')} \cdot \min_{H(\boldsymbol{\eta},\boldsymbol{\eta}')=1} \big\|Q_{f_{\boldsymbol{\eta}}} \wedge Q_{f_{\boldsymbol{\eta}'}}\big\|$$

for every estimator $\hat{f}_n$ based on $(y_1,\mathbf{x}^{(1)}),\ldots,(y_n,\mathbf{x}^{(n)})$, where Q_f is the joint distribution of $(y_i,\mathbf{x}^{(i)})$ in (3.25) when $f^* = f$. Here, $\|Q_{f_{\boldsymbol{\eta}}} \wedge Q_{f_{\boldsymbol{\eta}'}}\|$ is defined by

$$\big\|Q_{f_{\boldsymbol{\eta}}} \wedge Q_{f_{\boldsymbol{\eta}'}}\big\| = \int \min(q_{f_{\boldsymbol{\eta}}}, q_{f_{\boldsymbol{\eta}'}})\, d\mu,$$

where μ is any measure for which both $q_{f_{\boldsymbol{\eta}}} := dQ_{f_{\boldsymbol{\eta}}}/d\mu$ and $q_{f_{\boldsymbol{\eta}'}} := dQ_{f_{\boldsymbol{\eta}'}}/d\mu$ exist. However, we can derive Lemma 18 directly from [149, Lemma 24.3]. This is simply because

$$\mathfrak{M}^{d,s}_{n,V} = \inf_{\hat{f}_n} \sup_{\substack{f^* \in \mathcal{F}^{d,s}_{\infty-mars} \\ V_{mars}(f^*)\leq V}} \mathbb{E}_{f^*}\|\hat{f}_n - f^*\|^2_{p_0,2} \geq \inf_{\hat{f}_n} \max_{\boldsymbol{\eta}} \mathbb{E}_{f_{\boldsymbol{\eta}}}\|\hat{f}_n - f_{\boldsymbol{\eta}}\|^2_{p_0,2}$$

and

$$\|Q_{f_{\boldsymbol{\eta}}} \wedge Q_{f_{\boldsymbol{\eta}'}}\| = \int \min(q_{f_{\boldsymbol{\eta}}}, q_{f_{\boldsymbol{\eta}'}})\, d\mu = 1 - \frac{1}{2}\int |q_{f_{\boldsymbol{\eta}}} - q_{f_{\boldsymbol{\eta}'}}|\, d\mu$$
$$\geq 1 - \sqrt{\frac{1}{2}K(Q_{f_{\boldsymbol{\eta}}}, Q_{f_{\boldsymbol{\eta}'}})} = 1 - \sqrt{\frac{1}{2}\mathbb{E}\big[K(\mathbb{P}_{f_{\boldsymbol{\eta}}}, \mathbb{P}_{f_{\boldsymbol{\eta}'}})\big]},$$

Here, the inequality is due to Pinsker's inequality (see, e.g., [152, Lemma 15.2]).

Proof of Theorem 8. Let l be a positive integer whose value will be specified momentarily. Consider the set

$$P_l = \left\{(p_1, \dots, p_s) \in \mathbb{Z}^s_{\geq 0} : \sum_{j=1}^{s} p_j = l\right\}$$

and for each $\mathbf{p} = (p_1, \dots, p_s) \in P_l$, let

$$I_{\mathbf{p}} = \big\{(i_1, \dots, i_s) : i_j \in [2^{p_j}] \text{ for each } j \in [s]\big\}.$$

Clearly, $|I_{\mathbf{p}}| = 2^l$ for every $\mathbf{p} \in P_l$. Also, we have

$$|P_l| = \binom{s+l-1}{s-1} \geq \frac{l^{s-1}}{(s-1)!} =: a_s l^{s-1}.$$

Now, we choose l as

$$l = \left\lceil \frac{1}{5\log 2}\left\{\log\left(\frac{C_{B,s}nV^2}{\sigma^2}\right) - (s-1)\log\log\left(\frac{C_{B,s}nV^2}{\sigma^2}\right)\right\}\right\rceil,$$

where $C_{B,s} = B2^{-12s+1}(10\log 2)^{s-1}/a_s$. Here, $\lceil x \rceil$ indicates the least integer greater than or equal to x. The reason why we choose these particular $C_{B,s}$ and l will become clear at a later stage of the proof. Note that

$$2^{-l} \leq \left(\frac{\sigma^2}{C_{B,s}nV^2}\right)^{1/5}\left[\log\left(\frac{C_{B,s}nV^2}{\sigma^2}\right)\right]^{(s-1)/5} < 2^{-l+1}. \tag{C.27}$$

Let

$$Q = \big\{(\mathbf{p}, \mathbf{i}) : \mathbf{p} \in P_l \text{ and } \mathbf{i} \in I_{\mathbf{p}}\big\}$$

and let $q = |Q| = |P_l| \cdot 2^l$. Throughout this proof, we will index the components of every $\boldsymbol{\eta} \in \{-1, 1\}^q$ with the set Q. For a positive integer m and $k \in [2^m]$, we let $\psi_{m,k}$ be the real-valued function on $(0,1)$ defined by

$$\psi_{m,k}(x) = \begin{cases} 1 & \text{if } x \in ((k-1)2^{-m}, (k-7/8)2^{-m}) \cup ((k-5/8)2^{-m}, (k-1/2)2^{-m}) \\ & \qquad\qquad \cup((k-3/8)2^{-m}, (k-1/8)2^{-m}) \\ -1 & \text{if } x \in ((k-7/8)2^{-m}, (k-5/8)2^{-m}) \cup ((k-1/2)2^{-m}, (k-3/8)2^{-m}) \\ & \qquad\qquad \cup((k-1/8)2^{-m}, k2^{-m}) \\ 0 & \text{otherwise.} \end{cases}$$

We now construct our function $f_{\boldsymbol{\eta}}$ for each $\boldsymbol{\eta} \in \{-1, 1\}^q$ based on these functions $\psi_{m,k}$. For each $\boldsymbol{\eta} \in \{-1, 1\}^q$, we let $\nu_{\boldsymbol{\eta}}$ be the signed measure on $(0, 1)^s$ defined by

$$d\nu_{\boldsymbol{\eta}}(\mathbf{t}) = \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \left(\prod_{j=1}^{s} \psi_{p_j, i_j}(t_j) \right) d\mathbf{t},$$

and let $f_{\boldsymbol{\eta}}$ denote the real-valued function on $[0, 1]^d$ defined by

$$f_{\boldsymbol{\eta}}(x_1, \ldots, x_d) = \int_{(0,1)^s} \prod_{j=1}^{s} (x_j - t_j)_+ \, d\nu_{\boldsymbol{\eta}}(\mathbf{t}).$$

The following lemma, whose proof is given in Appendix C.22, then presents the key properties of the functions $f_{\boldsymbol{\eta}}$.

Lemma 19. *For each $\boldsymbol{\eta} \in \{-1, 1\}^q$, $f_{\boldsymbol{\eta}}$ satisfies*

$$V_{mars}(f_{\boldsymbol{\eta}}) \leq V. \tag{C.28}$$

Also, we have

$$\max_{H(\boldsymbol{\eta}, \boldsymbol{\eta}')=1} \|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2 \leq \frac{BV^2}{|P_l|} \cdot 2^{-5l-12s+2}. \tag{C.29}$$

and

$$\min_{\boldsymbol{\eta} \neq \boldsymbol{\eta}'} \frac{\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2}{H(\boldsymbol{\eta}, \boldsymbol{\eta}')} \geq \frac{bV^2}{|P_l|} \cdot 2^{-5l-14s+2}. \tag{C.30}$$

Since the Kullback divergence between $\mathbb{P}_{f_{\boldsymbol{\eta}}}$ and $\mathbb{P}_{f_{\boldsymbol{\eta}'}}$ for $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \{-1, 1\}^q$ is given by

$$K(\mathbb{P}_{f_{\boldsymbol{\eta}}}, \mathbb{P}_{f_{\boldsymbol{\eta}'}}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(f_{\boldsymbol{\eta}}(\mathbf{x}^{(i)}) - f_{\boldsymbol{\eta}'}(\mathbf{x}^{(i)}) \right)^2,$$

we can obtain from (C.29) that

$$\max_{H(\boldsymbol{\eta}, \boldsymbol{\eta}')=1} \mathbb{E}\big[K(\mathbb{P}_{f_{\boldsymbol{\eta}}}, \mathbb{P}_{f_{\boldsymbol{\eta}'}})\big] = \frac{n}{2\sigma^2} \cdot \max_{H(\boldsymbol{\eta}, \boldsymbol{\eta}')=1} \|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2 \leq \frac{BnV^2}{\sigma^2 |P_l|} \cdot 2^{-5l-12s+1}. \tag{C.31}$$

We are now ready to apply Lemma 18. By Lemma 18 together with (C.30) and (C.31), the minimax risk $\mathfrak{M}_{n,V}^{d,s}$ satisfies

$$\begin{aligned} \mathfrak{M}_{n,V}^{d,s} &\geq \frac{q}{8} \cdot \frac{bV^2}{|P_l|} \cdot 2^{-5l-14s+2} \left(1 - \sqrt{\frac{BnV^2}{\sigma^2 |P_l|} \cdot 2^{-5l-12s}} \right) \\ &= bV^2 2^{-4l-14s-1} \left(1 - \sqrt{\frac{1}{2(10\log 2)^{s-1}} \cdot \frac{C_{B,s} n V^2}{\sigma^2} \cdot \frac{2^{-5l}}{l^{s-1}}} \right). \end{aligned}$$

Recall that we let $C_{B,s} = B2^{-12s+1}(10\log 2)^{s-1}/a_s$. Our choice of l then implies that

$$
\begin{aligned}
\frac{1}{2(10\log 2)^{s-1}} \cdot \frac{C_{B,s}nV^2}{\sigma^2} \cdot \frac{2^{-5l}}{l^{s-1}} &\leq \frac{1}{2}\left[\frac{\log(C_{B,s}nV^2/\sigma^2)}{2\{\log(C_{B,s}nV^2/\sigma^2) - (s-1)\log\log(C_{B,s}nV^2/\sigma^2)\}}\right]^{s-1} \\
&\leq \frac{1}{2}\left[2\left(1-(s-1)\frac{\log\log(C_{B,s}nV^2/\sigma^2)}{\log(C_{B,s}nV^2/\sigma^2)}\right)\right]^{-(s-1)} \\
&\leq \frac{1}{2}\left[2\left(1-(s-1)\left\{\log\left(\frac{C_{B,s}nV^2}{\sigma^2}\right)\right\}^{-1/2}\right)\right]^{-(s-1)}.
\end{aligned}
$$

Here we use

$$
l \geq \frac{1}{5\log 2}\left\{\log\left(\frac{C_{B,s}nV^2}{\sigma^2}\right) - (s-1)\log\log\left(\frac{C_{B,s}nV^2}{\sigma^2}\right)\right\}
$$

and (C.27) for the first inequality, and the last one is due to $\log\log x/\log x \leq (\log x)^{-1/2}$, which holds for all $x > 1$. Thus, if

$$
n \geq \frac{e^{4s^2}}{C_{B,s}} \cdot \frac{\sigma^2}{V^2},
$$

we have

$$
\frac{1}{2(10\log 2)^{s-1}} \cdot \frac{C_{B,s}nV^2}{\sigma^2} \cdot \frac{2^{-5l}}{l^{s-1}} \leq \frac{1}{2} \cdot \left(1+\frac{1}{s}\right)^{-(s-1)} \leq \frac{1}{2},
$$

and therefore,

$$
\begin{aligned}
\mathfrak{M}_{n,V}^{d,s} &\geq \left(1-\sqrt{\frac{1}{2}}\right) \cdot bV^2 2^{-4l-14s-1} \\
&\geq \left(1-\sqrt{\frac{1}{2}}\right) \cdot bV^2 2^{-14s-5} \cdot \left(\frac{\sigma^2}{C_{B,s}nV^2}\right)^{4/5}\left[\log\left(\frac{C_{B,s}nV^2}{\sigma^2}\right)\right]^{4(s-1)/5} \\
&\geq C_{b,B,s}\left(\frac{\sigma^2 V^{1/2}}{n}\right)^{4/5}\left[\log\left(\frac{C_{B,s}nV^2}{\sigma^2}\right)\right]^{4(s-1)/5},
\end{aligned}
$$

where $C_{b,B,s} = (1-1/\sqrt{2}) \cdot b2^{-14s-5}C_{B,s}^{-4/5}$. Here, for the second inequality, we use (C.27) again. Now, note that

$$
\log\left(\frac{C_{B,s}nV^2}{\sigma^2}\right) \geq \frac{1}{2}\log\left(\frac{nV^2}{\sigma^2}\right)
$$

if $n \geq (1/C_{B,s}^2) \cdot (\sigma^2/V^2)$. Consequently, by further assuming that

$$
n \geq \max\left(\frac{e^{4s^2}}{C_{B,s}} \cdot \frac{\sigma^2}{V^2}, \frac{1}{C_{B,s}^2} \cdot \frac{\sigma^2}{V^2}\right),
$$

we can derive the conclusion

$$
\mathfrak{M}_{n,V}^{d,s} \geq C'_{b,B,s}\left(\frac{\sigma^2 V^{1/2}}{n}\right)^{4/5}\left[\log\left(\frac{nV^2}{\sigma^2}\right)\right]^{4(s-1)/5},
$$

where $C'_{b,B,s} = C_{b,B,s} \cdot 2^{-4(s-1)/5}$. $\square$

C.14 Proof of Lemma 9

Proof of Lemma 9. Assume that a real-valued function g defined on $[0,1]^m$ has finite $\mathrm{HK}_{\mathbf{0}}$ variation, is right-continuous, and is left-continuous at each point $\mathbf{x} \in [0,1]^m \setminus [0,1)^m$ with respect to all the j^{th} coordinates where $x_j = 1$. Since g has finite $\mathrm{HK}_{\mathbf{0}}$ variation and is right-continuous, by Theorem 1, there exists a finite signed measure $\bar{\nu}$ on $[0,1]^m$ such that

$$g(\mathbf{x}) = \bar{\nu}\bigg(\prod_{j=1}^{m}[0,x_j]\bigg)$$

for all $\mathbf{x} \in [0,1]^m$. Let μ be the finite signed measure on $[0,1]^m$ for which

$$\mu(E) = \bar{\nu}(E \cap [0,1)^m)$$

for all Borel sets $E \subseteq [0,1]^m$. Then, it follows that

$$g(\mathbf{x}) = \mu\bigg(\prod_{j=1}^{m}[0,x_j]\bigg)$$

for all $\mathbf{x} \in [0,1]^m$. Indeed, if $\mathbf{x} \in [0,1)^m$, then

$$g(\mathbf{x}) = \bar{\nu}\bigg(\prod_{j=1}^{m}[0,x_j]\bigg) = \bar{\nu}\bigg(\prod_{j=1}^{m}[0,x_j] \cap [0,1)^m\bigg) = \mu\bigg(\prod_{j=1}^{m}[0,x_j]\bigg).$$

Otherwise, without loss of generality, let $\mathbf{x} = (1,\dots,1,x_{k+1},\dots,x_m)$ where $x_{k+1},\dots,x_m < 1$ for some $k \in [m]$. Then, by the left-continuity of g,

$$\begin{aligned} g(\mathbf{x}) &= \lim_{y_1 \to 1-} \cdots \lim_{y_k \to 1-} g(y_1,\dots,y_k,x_{k+1},\dots,x_m) = \lim_{y_1 \to 1-} \cdots \lim_{y_k \to 1-} \bar{\nu}\bigg(\prod_{j=1}^{k}[0,y_j] \times \prod_{j=k+1}^{m}[0,x_j]\bigg) \\ &= \bar{\nu}\big([0,1)^k \times [0,x_{k+1}] \times \cdots \times [0,x_m]\big) = \bar{\nu}\bigg(\prod_{j=1}^{m}[0,x_j] \cap [0,1)^m\bigg) = \mu\bigg(\prod_{j=1}^{m}[0,x_j]\bigg), \end{aligned}$$

as well. We thus have the two signed measures $\bar{\nu}$ and μ satisfying (2.4), and the uniqueness in Theorem 1 implies that $\bar{\nu} = \mu$, i.e.,

$$\bar{\nu}(E) = \mu(E) = \bar{\nu}(E \cap [0,1)^m)$$

for all Borel sets $E \subseteq [0,1]^m$. Therefore, if we let ν be the finite signed measure on $[0,1)^m$ for which

$$\nu(E) = \bar{\nu}(E),$$

then for every $\mathbf{x} \in [0,1]^m$,

$$g(\mathbf{x}) = \bar{\nu}\bigg(\prod_{j=1}^{m}[0,x_j]\bigg) = \bar{\nu}\bigg(\prod_{j=1}^{m}[0,x_j] \cap [0,1)^m\bigg) = \nu\bigg(\prod_{j=1}^{m}[0,x_j] \cap [0,1)^m\bigg)$$

as desired. The uniqueness of such a signed measure directly follows from the uniqueness of the signed measure in Theorem 1.

Next, suppose a real-valued function g on $[0,1]^m$ is of the form (C.3) for some signed measure ν on $[0,1)^m$. Let $\bar{\nu}$ be the finite signed measure on $[0,1]^m$ for which

$$\bar{\nu}(E) = \nu(E \cap [0,1)^m)$$

for all Borel sets $E \subseteq [0,1]^m$. Then, clearly,

$$g(\mathbf{x}) = \nu\bigg(\prod_{j=1}^{m}[0,x_j] \cap [0,1)^m\bigg) = \bar{\nu}\bigg(\prod_{j=1}^{m}[0,x_j]\bigg)$$

for every $\mathbf{x} \in [0,1]^m$. By Theorem 1, g thus has finite $\mathrm{HK}_{\mathbf{0}}$ variation and

$$|\bar{\nu}|([0,1]^m) = \mathrm{HK}_{\mathbf{0}}(g) + |g(\mathbf{0})|.$$

Also, since $|g(\mathbf{0})| = |\nu(\{\mathbf{0}\})| = |\nu|(\{\mathbf{0}\})$ and $|\bar{\nu}|([0,1]^m) = |\nu|([0,1)^m)$, it follows that

$$\mathrm{HK}_{\mathbf{0}}(g) = |\nu|([0,1)^m \setminus \{\mathbf{0}\}).$$

□

C.15 Proof of Lemma 12

Proof of Lemma 12. Suppose $f = f_{c,\{\nu_S\}} \in D(V,t)$. Let $c_\emptyset = c$, and let $c_S = \nu_S(\{\mathbf{0}\})$ for each $S \subseteq [d]$ with $0 < |S| \leq s$. Then, we can write f as

$$\begin{aligned} f(x_1,\dots,x_d) &= c + \sum_{0<|S|\leq s} \int_{[0,1)^{|S|}} \prod_{j\in S}(x_j - t_j)_+ \, d\nu_S(t_j, j\in S) \\ &= \sum_{0\leq|S|\leq s} c_S \prod_{j\in S} x_j + \sum_{0<|S|\leq s} \int_{[0,1)^{|S|}\setminus\{\mathbf{0}\}} \prod_{j\in S}(x_j - t_j)_+ \, d\nu_S(t_j, j\in S) \end{aligned}$$

for $(x_1,\dots,x_d) \in [0,1]^d$. Note that

$$t^2 \geq \|f\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} f^2(\mathbf{x}^{(i)}) = \frac{1}{n}\sum_{\mathbf{i}\in I_0} f^2\big(v_{i_1}^{(1)},\dots,v_{i_d}^{(d)}\big)$$

where $I_0 := \{0,\dots,n_1-1\}\times\cdots\times\{0,\dots,n_d-1\}$. By Cauchy inequality, we have

$$f^2(x_1,\dots,x_d) \geq \frac{1}{2}\bigg[\sum_{0\leq|S|\leq s} c_S \prod_{j\in S} x_j\bigg]^2 - \bigg[\sum_{0<|S|\leq s} \int_{[0,1)^{|S|}\setminus\{\mathbf{0}\}} \prod_{j\in S}(x_j - t_j)_+ \, d\nu_S(t_j, j\in S)\bigg]^2$$

$$\geq \frac{1}{2}\bigg[\sum_{0\leq|S|\leq s} c_S \prod_{j\in S} x_j\bigg]^2 - \bigg[\sum_{0<|S|\leq s}\int_{[0,1)^{|S|}\setminus\{\mathbf{0}\}} d|\nu_S|(t_j, j\in S)\bigg]^2$$
$$\geq \frac{1}{2}\bigg[\sum_{0\leq|S|\leq s} c_S \prod_{j\in S} x_j\bigg]^2 - V^2$$

for every $(x_1,\dots,x_d)\in[0,1]^d$. Using this inequality, we bound c_S for each S first. Observe that $|c_\emptyset| = |f(\mathbf{0})| \leq \sqrt{n}t \leq V + \sqrt{n}t$. The above inequality then implies that

$$nt^2 \geq \sum_{\mathbf{i}\in I_0} f^2\big(v^{(1)}_{i_1},\dots,v^{(d)}_{i_d}\big) \geq \sum_{\substack{i_1\geq 0\\ i_2,\dots,i_d=0}} f^2\big(v^{(1)}_{i_1},\dots,v^{(d)}_{i_d}\big) \geq \sum_{i_1=0}^{n_1-1}\Big[\frac{1}{2}\big(c_{\{1\}}v^{(1)}_{i_1}+c_\emptyset\big)^2 - V^2\Big]$$
$$= \frac{1}{2}\sum_{i_1=0}^{n_1-1}\big(c_{\{1\}}v^{(1)}_{i_1}+c_\emptyset\big)^2 - n_1V^2 \geq \frac{1}{2}\sum_{i_1=0}^{n_1-1}\Big(\frac{1}{2}c^2_{\{1\}}\big(v^{(1)}_{i_1}\big)^2 - c^2_\emptyset\Big) - n_1V^2$$
$$\geq \frac{1}{4}c^2_{\{1\}}\sum_{i_1=0}^{n_1-1}\big(v^{(1)}_{i_1}\big)^2 - \frac{n_1}{2}(nt^2) - n_1V^2 \geq \frac{1}{4}c^2_{\{1\}}\sum_{i_1=0}^{n_1-1}\rho^2\Big(\frac{i_1}{n_1}\Big)^2 - \frac{n_1}{2}(nt^2) - n_1V^2$$
$$= \rho^2\cdot\frac{(n_1-1)(2n_1-1)}{24n_1}\cdot c^2_{\{1\}} - \frac{n_1}{2}(nt^2) - n_1V^2.$$

Here, the last inequality follows from

$$v^{(1)}_{i_1} = v^{(1)}_0 + \sum_{l_1=1}^{i_1}\big(v^{(1)}_{l_1} - v^{(1)}_{l_1-1}\big) \geq 0 + \sum_{l_1=1}^{i_1}\frac{\rho}{n_1} = \rho\cdot\frac{i_1}{n_1}$$

for $i_1 = 1,\dots,n_1-1$. Further applying the inequality $(x+y)^{1/2}\leq x^{1/2}+y^{1/2}$, we obtain

$$|c_{\{1\}}| \leq C_\rho(V+\sqrt{n}t).$$

By the same argument, it can be shown that

$$|c_S| \leq C_\rho(V+\sqrt{n}t)$$

for every $S\subseteq[d]$ with $|S|=1$. Also, since

$$nt^2 \geq \sum_{\mathbf{i}\in I_0} f^2\big(v^{(1)}_{i_1},\dots,v^{(d)}_{i_d}\big) \geq \sum_{\substack{i_1,i_2\geq 0\\ i_3,\dots,i_d=0}}\Big[\frac{1}{2}\big(c_{\{1,2\}}v^{(1)}_{i_1}v^{(2)}_{i_2} + c_{\{1\}}v^{(1)}_{i_1} + c_{\{2\}}v^{(2)}_{i_2} + c_\emptyset\big)^2 - V^2\Big]$$
$$= \frac{1}{2}\sum_{i_1=0}^{n_1-1}\sum_{i_2=0}^{n_2-1}\big(c_{\{1,2\}}v^{(1)}_{i_1}v^{(2)}_{i_2} + c_{\{1\}}v^{(1)}_{i_1} + c_{\{2\}}v^{(2)}_{i_2} + c_\emptyset\big)^2 - n_1n_2V^2$$
$$\geq \frac{1}{2}\sum_{i_1=0}^{n_1-1}\sum_{i_2=0}^{n_2-1}\Big[\frac{1}{2}c^2_{\{1,2\}}\big(v^{(1)}_{i_1}v^{(2)}_{i_2}\big)^2 - \big(c_{\{1\}}v^{(1)}_{i_1} + c_{\{2\}}v^{(2)}_{i_2} + c_\emptyset\big)^2\Big] - n_1n_2V^2$$

$$\geq \rho^4 \Big[\frac{(n_1-1)(2n_1-1)}{12n_1} \cdot \frac{(n_2-1)(2n_2-1)}{12n_2} \Big] \cdot c_{\{1,2\}}^2 - C_\rho n_1 n_2 (V^2 + nt^2),$$

it follows that

$$|c_{\{1,2\}}| \leq C_\rho (V + \sqrt{n} t).$$

The same argument implies

$$|c_S| \leq C_\rho (V + \sqrt{n} t)$$

for all $S \subseteq [d]$ with $|S| = 2$. Repeating this argument, we can show inductively that

$$|c_S| \leq C_{\rho,s} (V + \sqrt{n} t) =: \widetilde{t}$$

for all $S \subseteq [d]$ with $0 \leq |S| \leq s$.

Now, fix $\delta > 0$, which will be specified later, and let

$$\mathcal{K} = \big\{ (k_S, S \subseteq [d], 0 \leq |S| \leq s) : -(K+1) \leq k_S \leq K \text{ for all } S \subseteq [d] \text{ with } 0 \leq |S| \leq s \big\},$$

where $K = \lfloor \widetilde{t}/\delta \rfloor$. Here, $\lfloor \widetilde{t}/\delta \rfloor$ indicates the greatest integer less than or equal to $\widetilde{t}/\delta$. Also, for each $\mathbf{k} \in \mathcal{K}$, let

$$\mathcal{G}_{\mathbf{k}} = \big\{ f_{c,\{\nu_S\}} \in D(V,t) : k_S \delta \leq c_S \leq (k_S+1)\delta \text{ for each } S \subseteq [d] \text{ with } 0 \leq |S| \leq s \big\}.$$

Then,

$$D(V,t) = \bigcup_{\mathbf{k} \in \mathcal{K}} \mathcal{G}_{\mathbf{k}},$$

and thereby,

$$\log N(\epsilon, D(V,t), \|\cdot\|_n) \leq \log \Big(\sum_{\mathbf{k} \in \mathcal{K}} N(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_n) \Big) \tag{C.32}$$

$$\leq \log |\mathcal{K}| + \sup_{\mathbf{k} \in \mathcal{K}} \log N(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_n) \leq 2^d \log \Big(2 + \frac{2\widetilde{t}}{\delta} \Big) + \sup_{\mathbf{k} \in \mathcal{K}} \log N(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_n).$$

Fix $\mathbf{k} \in \mathcal{K}$. Let $\mathcal{G}_{\mathbf{k}}^{\emptyset}$ be the collection of all the constant functions on $[0,1]^d$

$$(x_1, \dots, x_d) \mapsto c$$

where $k_\emptyset \delta \leq c \leq (k_\emptyset + 1)\delta$, and for each $S \subseteq [d]$ with $0 < |S| \leq s$, let $\mathcal{G}_{\mathbf{k}}^S$ be the collection of the functions on $[0,1]^d$ of the form

$$(x_1, \dots, x_d) \mapsto c_S \prod_{j \in S} x_j + \int_{[0,1)^{|S|} \setminus \{\mathbf{0}\}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S),$$

where $k_S \delta \leq c_S \leq (k_S + 1)\delta$ and ν_S is a signed measure supported on $(\prod_{j \in S} \mathcal{V}_j) \cap [0,1)^{|S|}$ with $|\nu_S|([0,1)^{|S|} \setminus \{\mathbf{0}\}) \leq V$. Then, by definition,

$$\mathcal{G}_{\mathbf{k}} \subseteq +_{0 \leq |S| \leq s} \mathcal{G}_{\mathbf{k}}^S,$$

which implies that

$$\log N(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_n) \le \sum_{0 \le |S| \le s} \log N\Big(\frac{\epsilon}{2^d}, \mathcal{G}_{\mathbf{k}}^S, \|\cdot\|_n\Big). \tag{C.33}$$

Here, for a collection of sets $\{X_i\}_{i \in I}$, $+_{i \in I} X_i$ indicates the set $\{\sum_{i \in I} x_i : x_i \in X_i \text{ for } i \in I\}$.

Now, we will bound each term of the right-hand side of (C.33). It is easy to show that

$$N(\epsilon, \mathcal{G}_{\mathbf{k}}^{\emptyset}, \|\cdot\|_n) \le 1 + \frac{\delta}{\epsilon}. \tag{C.34}$$

Hence, it suffices to bound $\log N(\epsilon, \mathcal{G}_{\mathbf{k}}^S, \|\cdot\|_n)$ for each $S \subseteq [d]$ with $0 < |S| \le s$. Fix $S \subseteq [d]$ with $0 < |S| \le s$, and let $\widetilde{\mathcal{G}}_{\mathbf{k}}^S$ be the collection of all functions on $[0,1]^{|S|}$ of the form

$$(x_j, j \in S) \mapsto c_S \prod_{j \in S} x_j + \int_{[0,1)^{|S|} \setminus \{\mathbf{0}\}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S),$$

where c_S and ν_S satisfy the same conditions as in the definition of $\mathcal{G}_{\mathbf{k}}^S$. With slight abuse of notation, let $\|\cdot\|_n$ also denote the norm on $[0,1]^{|S|}$ defined by

$$\|f\|_n^2 = \frac{1}{\prod_{j \in S} n_j} \sum_{\mathbf{i} \in \prod_{j \in S} \{0, \ldots, n_j - 1\}} f^2\big(v_{i_j}^{(j)}, j \in S\big).$$

Then, clearly,

$$\log N(\epsilon, \mathcal{G}_{\mathbf{k}}^S, \|\cdot\|_n) \le \log N(\epsilon, \widetilde{\mathcal{G}}_{\mathbf{k}}^S, \|\cdot\|_n). \tag{C.35}$$

Also, since $\widetilde{\mathcal{G}}_{\mathbf{k}}^S$ can be obtained by shifting $\widetilde{\mathcal{G}}_{\mathbf{0}}^S$ by some suitable function, we have

$$\log N(\epsilon, \widetilde{\mathcal{G}}_{\mathbf{k}}^S, \|\cdot\|_n) = \log N(\epsilon, \widetilde{\mathcal{G}}_{\mathbf{0}}^S, \|\cdot\|_n). \tag{C.36}$$

Moreover, if we let $\widetilde{\mathcal{G}}_S$ be the collection of all functions on $[0,1]^{|S|}$ of the form

$$(x_j, j \in S) \mapsto \int_{[0,1)^{|S|}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S)$$

where ν_S is a signed measure supported on $(\prod_{j \in S} \mathcal{V}_j) \cap [0,1)^{|S|}$ with $|\nu_S|([0,1)^{|S|}) \le V + \delta$, then

$$\log N(\epsilon, \widetilde{\mathcal{G}}_{\mathbf{0}}^S, \|\cdot\|_n) \le \log N(\epsilon, \widetilde{\mathcal{G}}_S, \|\cdot\|_n) \tag{C.37}$$

because $\widetilde{\mathcal{G}}_{\mathbf{0}}^S \subseteq \widetilde{\mathcal{G}}_S$. For these reasons, we can see that it is enough to bound the metric entropy of $\widetilde{\mathcal{G}}_S$ in order to bound the metric entropy of $\mathcal{G}_{\mathbf{k}}^S$.

For $m \in [d]$ and $R > 0$, we denote by $\mathcal{D}_{m,\mathrm{disc}}(R)$ the collection of the functions on $[0,1]^m$ of the form

$$(x_1, \ldots, x_m) \mapsto \int_{[0,1)^m} (x_1 - t_1)_+ \cdots (x_m - t_m)_+ \, d\nu(\mathbf{t})$$

where ν is a signed measure supported on $(\prod_{j=1}^m \mathcal{V}_j) \cap [0,1)^m$ with total variation $|\nu|([0,1)^m) \leq R$. We can see that

$$\widetilde{\mathcal{G}}_S = \mathcal{D}_{|S|,\mathrm{disc}}(V+\delta)$$

after suitable re-indexing. From the following lemma, which provides an upper bound on the metric entropy of $\mathcal{D}_{|S|,\mathrm{disc}}(R)$, we can therefore obtain an upper bound on the metric entropy of $\widetilde{\mathcal{G}}_S$. We defer our proof of this lemma to Appendix C.23.

Lemma 20. *There exist positive constants $C_{\rho,m}$ and $\kappa_{\rho,m}$ depending on ρ and m such that for every $R > 0$,*

$$\log N(\epsilon, \mathcal{D}_{m,disc}(R), \|\cdot\|_n) \leq C_{\rho,m}\Big(\frac{R}{\epsilon}\Big)^{1/2}\Big[\log\Big(\frac{R}{\epsilon}\Big)\Big]^{3(2m-1)/4}$$

provided that $0 < \epsilon/R \leq \kappa_{\rho,m}$. The log term can be omitted when $m = 1$.

By Lemma 20,

$$\begin{aligned}\log N(\epsilon, \widetilde{\mathcal{G}}_S, \|\cdot\|_n) &\leq C_{\rho,|S|}\Big(\frac{V+\delta}{\epsilon}\Big)^{1/2}\Big[\log\Big(\frac{V+\delta}{\epsilon}\Big)\Big]^{3(2|S|-1)/4} \\ &\leq C_{\rho,|S|}\Big(\frac{2(V+\delta)}{\epsilon}\Big)^{1/2}\Big[\log\Big(\frac{2(V+\delta)}{\epsilon}\Big)\Big]^{3(2|S|-1)/4} \qquad \text{(C.38)}\end{aligned}$$

provided that $0 < \epsilon \leq \kappa_{\rho,|S|}(V+\delta) =: \epsilon_0$. Note that the logarithmic multiplicative factor can be omitted when $|S| = 1$. For $\epsilon_0 < \epsilon \leq V + \delta$,

$$\log N(\epsilon, \widetilde{\mathcal{G}}_S, \|\cdot\|_n) \leq \log N(\epsilon_0, \widetilde{\mathcal{G}}_S, \|\cdot\|_n) \leq \widetilde{C}_{\rho,|S|}\Big(\frac{2(V+\delta)}{\epsilon}\Big)^{1/2}\Big[\log\Big(\frac{2(V+\delta)}{\epsilon}\Big)\Big]^{3(2|S|-1)/4}$$

for

$$\widetilde{C}_{\rho,|S|} = C_{\rho,|S|}\Big(\frac{1}{\kappa_{\rho,|S|}}\Big)^{1/2}\Big[\log\Big(\frac{2}{\kappa_{\rho,|S|}}\Big)\Big]^{3(2|S|-1)/4}\big[2^{1/2}(\log 2)^{3(2|S|-1)/4}\big]^{-1},$$

where $C_{\rho,|S|}$ and $\kappa_{\rho,|S|}$ are the constants in (C.38). Moreover, for $f \in \widetilde{\mathcal{G}}_S$, we have

$$\|f\|_n \leq V + \delta,$$

which implies that $\log N(\epsilon, \widetilde{\mathcal{G}}_S, \|\cdot\|_n) = 0$ for every $\epsilon > V + \delta$. Combining those pieces, we can deduce that

$$\log N(\epsilon, \widetilde{\mathcal{G}}_S, \|\cdot\|_n) \leq C_{\rho,|S|}\Big(1 + \frac{2(V+\delta)}{\epsilon}\Big)^{1/2}\Big[\log\Big(1 + \frac{2(V+\delta)}{\epsilon}\Big)\Big]^{3(2|S|-1)/4}$$

for every $\epsilon > 0$. This together with (C.35), (C.36), and (C.37) implies that

$$\log N(\epsilon, \mathcal{G}^S_{\mathbf{k}}, \|\cdot\|_n) \leq C_{\rho,|S|}\Big(1 + \frac{2(V+\delta)}{\epsilon}\Big)^{1/2}\Big[\log\Big(1 + \frac{2(V+\delta)}{\epsilon}\Big)\Big]^{3(2|S|-1)/4} \qquad \text{(C.39)}$$

Note that we can omit the logarithmic multiplicative factor when $|S| = 1$.

By (C.33), (C.34), and (C.39), we have

$$\log N(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_n) \le \log\Big(1+\frac{2^d\delta}{\epsilon}\Big) + C_\rho \sum_{|S|=1}\Big[1+\frac{2^{d+1}(V+\delta)}{\epsilon}\Big]^{1/2} + C_{\rho,d}\sum_{1<|S|\le s}\Big[1+\frac{2^{d+1}(V+\delta)}{\epsilon}\Big]^{1/2}\Big[\log\Big(1+\frac{2^{d+1}(V+\delta)}{\epsilon}\Big)\Big]^{3(2|S|-1)/4}.$$

As a result, with the choice of $\delta=\epsilon$, (C.32) yields the conclusion

$$\begin{aligned}\log N(\epsilon, D(V,t), \|\cdot\|_n) &\le 2^d \log\Big(2 + C_{\rho,s}\cdot\frac{V+\sqrt{n}t}{\epsilon}\Big) + C_{\rho,d} + C_{\rho,d}\Big(\frac{V}{\epsilon}\Big)^{1/2} \\ &\quad + C_{\rho,d}\Big(2^{d+1}+1+\frac{2^{d+1}V}{\epsilon}\Big)^{1/2}\Big[\log\Big(2^{d+1}+1+\frac{2^{d+1}V}{\epsilon}\Big)\Big]^{3(2s-1)/4} \\ &\le 2^d \log\Big(2 + C_{\rho,s}\cdot\frac{V+\sqrt{n}t}{\epsilon}\Big) \\ &\quad + C_{\rho,d}\Big(2^{d+1}+1+\frac{2^{d+1}V}{\epsilon}\Big)^{1/2}\Big[\log\Big(2^{d+1}+1+\frac{2^{d+1}V}{\epsilon}\Big)\Big]^{3(2s-1)/4}.\end{aligned}$$

□

C.16 Proof of Lemma 13

Proof of Lemma 13. We basically follow the proof of [46, Lemma C.5] here. First, note that

$$\int_0^t \Big(\frac{u}{\epsilon}\Big)^{1/4}\Big[\log\frac{u}{\epsilon}\Big]^k d\epsilon = u\int_\tau^\infty e^{-3v/4}v^k\,dv,$$

where $v=\log(u/\epsilon)$ and $\tau=\log(u/t)$. If $\tau\le 1$,

$$\int_0^t \Big(\frac{u}{\epsilon}\Big)^{1/4}\Big[\log\frac{u}{\epsilon}\Big]^k d\epsilon \le u\int_0^\infty e^{-3v/4}v^k\,dv \le Cue^{-3\tau/4} = Cu^{1/4}t^{3/4},$$

so the lemma directly follows. Otherwise, applying integration by parts iteratively, we obtain

$$\begin{aligned}\int_\tau^\infty e^{-3v/4}v^k\,dv &= \frac{4}{3}e^{-3\tau/4}\tau^k + \frac{4}{3}k\int_\tau^\infty e^{-3v/4}v^{k-1}\,dv \\ &= \cdots = \frac{4}{3}e^{-3\tau/4}\tau^k + \cdots + \Big(\frac{4}{3}\Big)^{\lfloor k\rfloor+1}k\cdots(k-\lfloor k\rfloor+1)e^{-3\tau/4}\tau^{k-\lfloor k\rfloor} \\ &\quad + \Big(\frac{4}{3}\Big)^{\lfloor k\rfloor+1}k\cdots(k-\lfloor k\rfloor)\int_\tau^\infty e^{-3v/4}v^{k-\lfloor k\rfloor-1}\,dv \\ &\le C_k e^{-3\tau/4}\tau^k + C_k\int_\tau^\infty e^{-3v/4}v^{k-\lfloor k\rfloor-1}\,dv\end{aligned}$$

$$\leq C_k e^{-3\tau/4}\tau^k + C_k e^{-3\tau/4} = C_k u^{-3/4} t^{3/4}(1+\tau^k).$$

Here, $\lfloor k \rfloor$ indicates the greatest integer less than or equal to k, and the inequalities are due to the fact that $\tau > 1$. From this result, we can derive the conclusion

$$\int_0^t \left(\frac{u}{\epsilon}\right)^{1/4} \left[\log \frac{u}{\epsilon}\right]^k d\epsilon \leq C_k u^{1/4} t^{3/4}(1+\tau^k).$$

□

C.17 Proof of Lemma 14

Proof of Lemma 14. For $S \subseteq [d]$ with $0 < |S| \leq s$ and $\mathbf{l} = (l_j, j \in S) \in \prod_{j\in S}\{0,\dots,N_j - 1\}$, let $R_{\mathbf{l}}^S = \prod_{j \in S}[l_j/N_j, (l_j+1)/N_j)$ and define discrete signed measures $\widetilde{\mu}_{S,\mathbf{l}}$ and μ_S as in the proof of Lemma 2. We will show that $f_{c,\{\mu_S\}}$ constructed from these signed measures μ_S satisfies all the desired conditions.

First, as in the proof of Lemma 2, it can be shown that

$$f_{c,\{\mu_S\}}\Big(\frac{i_1}{N_1},\dots,\frac{i_d}{N_d}\Big) = f_{c,\{\nu_S\}}\Big(\frac{i_1}{N_1},\dots,\frac{i_d}{N_d}\Big) \tag{C.40}$$

for every $(i_1,\dots,i_d) \in \prod_{j=1}^d \{0,\dots,N_j\}$ and

$$V_{\mathrm{mars}}(f_{c,\{\mu_S\}}) \leq V_{\mathrm{mars}}(f_{c,\{\nu_S\}}).$$

Also, by repeating the arguments in the proof of Lemma 2, we can show that

$$\int_{[0,1]^{|S|}} \prod_{j\in S}(x_j - t_j)_+ \, d\widetilde{\mu}_{S,\mathbf{l}}(t_j, j\in S) = \int_{R_{\mathbf{l}}^S} \prod_{j\in S}(x_j - t_j)_+ \, d\nu_S(t_j, j \in S) \tag{C.41}$$

if $x_j \geq (l_j+1)/N_j$ for all $j \in S$, and that

$$|\widetilde{\mu}_{S,\mathbf{l}}|([0,1]^{|S|}) \leq |\nu_S|\big(R_{\mathbf{l}}^S\big). \tag{C.42}$$

Fix $(x_1,\dots,x_d) \in [0,1)^d$, and suppose that $i_j/N_j \leq x_j < (i_j+1)/N_j$ for $j \in [d]$. Then,

$$\begin{aligned}
&\big|f_{c,\{\mu_S\}}(x_1,\dots,x_d) - f_{c,\{\nu_S\}}(x_1,\dots,x_d)\big| \\
&\quad \leq \sum_{0<|S|\leq s} \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\dots,i_j\}} \bigg| \int_{[0,1]^{|S|}} \prod_{j\in S}(x_j - t_j)_+ \, d\widetilde{\mu}_{S,\mathbf{l}}(t_j, j\in S) \\
&\qquad\qquad\qquad - \int_{R_{\mathbf{l}}^S} \prod_{j\in S}(x_j - t_j)_+ \, d\nu_S(t_j, j\in S)\bigg|.
\end{aligned}$$

If $\mathbf{l} \in \prod_{j\in S}\{0,\ldots,i_j-1\}$, then

$$\left|\int_{[0,1]^{|S|}} \prod_{j\in S}(x_j-t_j)_+ \, d\widetilde{\mu}_{S,\mathbf{l}}(t_j, j\in S) - \int_{R^S_{\mathbf{l}}} \prod_{j\in S}(x_j-t_j)_+ \, d\nu_S(t_j, j\in S)\right| = 0$$

because of (C.41). Otherwise, if $\mathbf{l} \in \prod_{j\in S}\{0,\ldots,i_j\}\setminus\prod_{j\in S}\{0,\ldots,i_j-1\}$, there exists $j_0\in S$ such that $l_{j_0} = i_{j_0}$. Hence, by (C.42),

$$\begin{aligned}
&\left|\int_{[0,1]^{|S|}} \prod_{j\in S}(x_j-t_j)_+ \, d\widetilde{\mu}_{S,\mathbf{l}}(t_j, j\in S) - \int_{R^S_{\mathbf{l}}} \prod_{j\in S}(x_j-t_j)_+ \, d\nu_S(t_j, j\in S)\right| \\
&\quad \le \frac{1}{N_{j_0}}\cdot\Big(|\widetilde{\mu}_{S,\mathbf{l}}|\big([0,1]^{|S|}\big) + |\nu_S|\big(R^S_{\mathbf{l}}\big)\Big) \le \frac{2}{N}\cdot|\nu_S|\big(R^S_{\mathbf{l}}\big) = \frac{2}{N}\cdot|\nu_S|\big(R^S_{\mathbf{l}}\setminus\{\mathbf{0}\}\big)
\end{aligned}$$

if $\mathbf{l}\neq\mathbf{0}$, and

$$\begin{aligned}
&\left|\int_{[0,1]^{|S|}} \prod_{j\in S}(x_j-t_j)_+ \, d\widetilde{\mu}_{S,\mathbf{0}}(t_j, j\in S) - \int_{R^S_{\mathbf{0}}} \prod_{j\in S}(x_j-t_j)_+ \, d\nu_S(t_j, j\in S)\right| \\
&\le \left|\int_{[0,1]^{|S|}\setminus\{\mathbf{0}\}} \prod_{j\in S}(x_j-t_j)_+ \, d\widetilde{\mu}_{S,\mathbf{0}}(t_j, j\in S)\right| + \left|\int_{R^S_{\mathbf{0}}\setminus\{\mathbf{0}\}} \prod_{j\in S}(x_j-t_j)_+ \, d\nu_S(t_j, j\in S)\right| \\
&\quad + \left|\prod_{j\in S} x_j \cdot \big(|\widetilde{\mu}_{S,\mathbf{0}}|(\{\mathbf{0}\}) - |\nu_S|(\{\mathbf{0}\})\big)\right| \\
&\le \frac{1}{N_{j_0}}\cdot\Big(|\widetilde{\mu}_{S,\mathbf{0}}|\big([0,1]^{|S|}\setminus\{\mathbf{0}\}\big) + |\nu_S|\big(R^S_{\mathbf{0}}\setminus\{\mathbf{0}\}\big) + \big||\widetilde{\mu}_{S,\mathbf{0}}|(\{\mathbf{0}\}) - |\nu_S|(\{\mathbf{0}\})\big|\Big) \\
&\le \frac{1}{N}\cdot\Bigg(|\nu_S|\big(R^S_{\mathbf{0}}\setminus\{\mathbf{0}\}\big) \\
&\qquad + \sum_{\boldsymbol{\delta}\in\{0,1\}^{|S|}\setminus\{\mathbf{0}\}} \left|\int_{R^S_{\mathbf{0}}\setminus\{\mathbf{0}\}} \prod_{j\in S}\left[\left(\frac{1/N_j - t_j}{1/N_j - 0}\right)^{1-\delta_j}\left(\frac{t_j-0}{1/N_j-0}\right)^{\delta_j}\right] d\nu_S(t_j, j\in S)\right| \\
&\qquad + \left|\int_{R^S_{\mathbf{0}}\setminus\{\mathbf{0}\}} \prod_{j\in S}\left(\frac{1/N_j - t_j}{1/N_j - 0}\right) d\nu_S(t_j, j\in S)\right|\Bigg) \\
&\le \frac{1}{N}\cdot\Bigg(|\nu_S|\big(R^S_{\mathbf{0}}\setminus\{\mathbf{0}\}\big) \\
&\qquad + \sum_{\boldsymbol{\delta}\in\{0,1\}^{|S|}} \left|\int_{R^S_{\mathbf{0}}\setminus\{\mathbf{0}\}} \prod_{j\in S}\left[\left(\frac{1/N_j - t_j}{1/N_j - 0}\right)^{1-\delta_j}\left(\frac{t_j-0}{1/N_j-0}\right)^{\delta_j}\right] d\nu_S(t_j, j\in S)\right|\Bigg) \\
&\le \frac{2}{N}\cdot|\nu_S|\big(R^S_{\mathbf{0}}\setminus\{\mathbf{0}\}\big),
\end{aligned}$$

where the third inequality is by the definition of $\widetilde{\mu}_{S,\mathbf{0}}$. Using these results, we first show that

$$\big|f_{c,\{\mu_S\}}(x_1,\ldots,x_d) - f_{c,\{\nu_S\}}(x_1,\ldots,x_d)\big|$$

$$
\begin{aligned}
&\leq \sum_{0<|S|\leq s} \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\ldots,i_j\}} \bigg| \int_{[0,1]^{|S|}} \prod_{j\in S} (x_j - t_j)_+ \, d\widetilde{\mu}_{S,\mathbf{l}}(t_j, j\in S) \\
&\qquad\qquad\qquad - \int_{R_{\mathbf{l}}^S} \prod_{j\in S} (x_j - t_j)_+ \, d\nu_S(t_j, j\in S) \bigg| \\
&\leq \sum_{0<|S|\leq s} \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\ldots,i_j\}} \frac{2}{N} \cdot |\nu_S|\big(R_{\mathbf{l}}^S \setminus \{\mathbf{0}\}\big) \\
&\leq \frac{2}{N} \sum_{0<|S|\leq s} |\nu_S|\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big) = \frac{2}{N} \cdot V_{\mathrm{mars}}(f_{c,\{\nu_S\}}),
\end{aligned}
$$

which directly leads to

$$
\|f_{c,\{\mu_S\}} - f_{c,\{\nu_S\}}\|_\infty \leq \frac{2}{N} \cdot V_{\mathrm{mars}}(f_{c,\{\nu_S\}}).
$$

Also, we can show that

$$
\begin{aligned}
&\big|f_{c,\{\mu_S\}}(x_1,\ldots,x_d) - f_{c,\{\nu_S\}}(x_1,\ldots,x_d)\big| \\
&\quad\leq \sum_{0<|S|\leq s} \sum_{j_0\in S} \sum_{\substack{\mathbf{l}\in\prod_{j\in S}\{0,\ldots,i_j\} \\ l_{j_0}=i_{j_0}}} \bigg| \int_{[0,1]^{|S|}} \prod_{j\in S} (x_j - t_j)_+ \, d\widetilde{\mu}_{S,\mathbf{l}}(t_j, j\in S) \\
&\qquad\qquad\qquad - \int_{R_{\mathbf{l}}^S} \prod_{j\in S} (x_j - t_j)_+ \, d\nu_S(t_j, j\in S) \bigg| \\
&\quad\leq \sum_{0<|S|\leq s} \sum_{j_0\in S} \sum_{\substack{\mathbf{l}\in\prod_{j\in S}\{0,\ldots,N_j-1\} \\ l_{j_0}=i_{j_0}}} \frac{2}{N} \cdot |\nu_S|\big(R_{\mathbf{l}}^S \setminus \{\mathbf{0}\}\big) \\
&\quad\leq \frac{2}{N} \cdot \sum_{0<|S|\leq s} \sum_{j_0\in S} |\nu_S|\Bigg(\bigg(\Big[\frac{i_{j_0}}{N_{j_0}}, \frac{i_{j_0}+1}{N_{j_0}}\Big) \times \prod_{\substack{j\in S \\ j\neq j_0}} [0,1)\bigg) \setminus \{\mathbf{0}\}\Bigg).
\end{aligned}
$$

From this result, we can derive that

$$
\begin{aligned}
&\|f_{c,\{\nu_S\}} - f_{c,\{\mu_S\}}\|_{p_0,2}^2 = \int_{[0,1]^d} \big(f_{c,\{\mu_S\}}(x_1,\ldots,x_d) - f_{c,\{\nu_S\}}(x_1,\ldots,x_d)\big)^2 p_0(\mathbf{x}) d\mathbf{x} \\
&\quad\leq B \cdot \sum_{\mathbf{i}\in\prod_{j=1}^d\{0,\ldots,N_j-1\}} \prod_{j=1}^d \frac{1}{N_j} \\
&\qquad\qquad \cdot \Bigg[\frac{2}{N} \sum_{0<|S|\leq s} \sum_{j_0\in S} |\nu_S|\Bigg(\bigg(\Big[\frac{i_{j_0}}{N_{j_0}}, \frac{i_{j_0}+1}{N_{j_0}}\Big) \times \prod_{\substack{j\in S \\ j\neq j_0}} [0,1)\bigg) \setminus \{\mathbf{0}\}\Bigg)\Bigg]^2
\end{aligned}
$$

$$
\begin{aligned}
&\le \frac{4B}{N^2} \cdot \sum_{\mathbf{i}\in\prod_{j=1}^d\{0,\ldots,N_j-1\}} \prod_{j=1}^d \frac{1}{N_j} \cdot \Big(\sum_{0<|S|\le s} |S| \Big) \\
&\qquad\qquad \cdot \sum_{0<|S|\le s} \sum_{j_0\in S} \Big[|\nu_S| \Big(\Big(\Big[\frac{i_{j_0}}{N_{j_0}}, \frac{i_{j_0}+1}{N_{j_0}} \Big) \times \prod_{\substack{j\in S\\ j\neq j_0}} [0,1) \Big) \setminus \{\mathbf{0}\} \Big) \Big]^2 \\
&\le \frac{C_d B}{N^2} \cdot \sum_{0<|S|\le s} \sum_{j_0\in S} \sum_{\mathbf{i}\in\prod_{j=1}^d\{0,\ldots,N_j-1\}} \prod_{j=1}^d \frac{1}{N_j} \cdot \Big[|\nu_S| \Big(\Big(\Big[\frac{i_{j_0}}{N_{j_0}}, \frac{i_{j_0}+1}{N_{j_0}} \Big) \times \prod_{\substack{j\in S\\ j\neq j_0}} [0,1) \Big) \setminus \{\mathbf{0}\} \Big) \Big]^2 \\
&\le \frac{C_d B}{N^3} \cdot \sum_{0<|S|\le s} \sum_{j_0\in S} \sum_{i_{j_0}=0}^{N_{j_0}-1} \Big[|\nu_S| \Big(\Big(\Big[\frac{i_{j_0}}{N_{j_0}}, \frac{i_{j_0}+1}{N_{j_0}} \Big) \times \prod_{\substack{j\in S\\ j\neq j_0}} [0,1) \Big) \setminus \{\mathbf{0}\} \Big) \Big]^2 \\
&\le \frac{C_d B}{N^3} \cdot \sum_{0<|S|\le s} \sum_{j_0\in S} \Big[\sum_{i_{j_0}=0}^{N_{j_0}-1} |\nu_S| \Big(\Big(\Big[\frac{i_{j_0}}{N_{j_0}}, \frac{i_{j_0}+1}{N_{j_0}} \Big) \times \prod_{\substack{j\in S\\ j\neq j_0}} [0,1) \Big) \setminus \{\mathbf{0}\} \Big) \Big]^2 \\
&\le \frac{C_d B}{N^3} \cdot \sum_{0<|S|\le s} |S| \cdot \Big[|\nu_S| \big([0,1)^{|S|} \setminus \{\mathbf{0}\} \big) \Big]^2 \\
&\le \frac{C_d B}{N^3} \cdot \Big[\sum_{0<|S|\le s} |\nu_S| \big([0,1)^{|S|} \setminus \{\mathbf{0}\} \big) \Big]^2 = \frac{C_d B}{N^3} \cdot \big(V_{\mathrm{mars}}(f_{c,\{\nu_S\}}) \big)^2,
\end{aligned}
$$

where the second inequality is from Cauchy inequality. □

C.18 Proof of Theorem 24

Here, we almost follow the proof of [69, Proposition 2] and make a few modifications to the proof. As in the proof of [69, Proposition 2], we also use the standard peeling argument along with the following moment inequality for multiplier empirical processes.

Lemma 21 (Proposition 3.1 of [61]). *Let $\mathcal{F}$ be a countable collection of real-valued functions defined on $\mathcal{X}$. Suppose $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are i.i.d. random variables with law P on $\mathcal{X}$, and $\xi_1, \ldots, \xi_n$ are independent mean-zero random variables independent of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$. Then, there exists a universal positive constant C such that for every $p \ge 1$,*

$$
\begin{aligned}
\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)}) \Big|^p \Big] \le C^p \Big[\Big(\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)}) \Big| \Big] \Big)^p \\
+ p^{p/2} n^{p/2} \Big(\sup_{f\in\mathcal{F}} \|f\|_{P,2} \Big)^p \cdot \max_i \|\xi_i\|_2^p + p^p \mathbb{E}\Big[\max_i \Big(|\xi_i|^p \cdot \sup_{f\in\mathcal{F}} |f(\mathbf{x}^{(i)})|^p \Big) \Big] \Big].
\end{aligned}
$$

Proof of Theorem 24. Let A be the event where $\hat{f}$ is a least squares estimator over $\mathcal{F}$. By the assumption, we have $\mathbb{P}(A) \geq 1 - \epsilon$. Fix $r \geq 1$ and let

$$\mathcal{F}_j = \{f \in \mathcal{F} : 2^{j-1} r t_n < \|f - f_0\|_{P,2} \leq 2^j r t_n\}$$

for each positive integer j. Then, clearly,

$$\{f \in \mathcal{F} : \|f - f_0\|_{P,2} > r t_n\} = \biguplus_{j=1}^{\infty} \mathcal{F}_j,$$

and thus

$$\mathbb{P}(\|\hat{f} - f_0\|_{P,2} > r t_n) \leq \mathbb{P}(A^c) + \mathbb{P}(A \text{ and } \|\hat{f} - f_0\|_{P,2} > r t_n) \leq \epsilon + \sum_{j=1}^{\infty} \mathbb{P}(A \text{ and } \hat{f} \in \mathcal{F}_j). \quad \text{(C.43)}$$

Recall that $\biguplus$ denotes disjoint union. Let $(M_n(f) : f \in \mathcal{F})$ denote the stochastic process defined by

$$M_n(f) = \frac{2}{n} \sum_{i=1}^{n} \xi_i (f - f^*)(\mathbf{x}^{(i)}) - \frac{1}{n} \sum_{i=1}^{n} (f - f^*)^2(\mathbf{x}^{(i)})$$

and let $(M(f) : f \in \mathcal{F})$ denote the deterministic process defined by

$$M(f) = \mathbb{E}[M_n(f)] = -\|f - f^*\|_{P,2}^2.$$

Note that $M_n(f)$ can be alternatively represented as

$$M_n(f) = -\frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}^{(i)})\right)^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i^2.$$

Because $f_0 \in \mathcal{F}$, if $\hat{f}$ is a least squares estimator over $\mathcal{F}$, then it follows that

$$M_n(\hat{f}) - M_n(f_0) \geq 0.$$

Now, observe that

$$\begin{aligned} M_n(f) - M_n(f_0) &= \frac{2}{n} \sum_{i=1}^{n} \xi_i (f - f_0)(\mathbf{x}^{(i)}) - \frac{1}{n} \sum_{i=1}^{n} (f - f_0)^2(\mathbf{x}^{(i)}) \\ &\quad - \frac{2}{n} \sum_{i=1}^{n} (f - f_0)(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)}) \end{aligned}$$

and that

$$M(f) - M(f_0) = -\|f - f^*\|_{P,2}^2 + \|f_0 - f^*\|_{P,2}^2$$

$$= -\|f - f_0\|_{P,2}^2 - 2\mathbb{E}_{\mathbf{X}\sim P}\big[(f - f_0)(\mathbf{X}) \cdot (f_0 - f^*)(\mathbf{X})\big],$$

where $\mathbf{X}$ in the expectation is independent of all $\mathbf{x}^{(i)}$. By the assumption that $\|f_0 - f^*\|_{P,2} \leq t_n/4$, we thus have that for every $f \in \mathcal{F}_j$

$$\begin{aligned}
-\left(M(f) - M(f_0)\right) &= \|f - f_0\|_{P,2}^2 + 2\mathbb{E}_{\mathbf{X}\sim P}\big[(f - f_0)(\mathbf{X}) \cdot (f_0 - f^*)(\mathbf{X})\big] \\
&\geq \|f - f_0\|_{P,2} \cdot \big(\|f - f_0\|_{P,2} - 2\|f_0 - f^*\|_{P,2}\big) \geq 2^{j-1} r t_n \Big(2^{j-1} r t_n - \frac{t_n}{2}\Big) \geq 2^{2j-3} r^2 t_n^2,
\end{aligned}$$

where the first inequality is from Cauchy inequality. Hence, for each positive integer j,

$$\begin{aligned}
\mathbb{P}(A \text{ and } \hat{f} \in \mathcal{F}_j) &\leq \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \big(M_n(f) - M_n(f_0)\big) \geq 0\Big) \\
&\leq \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \big((M_n(f) - M_n(f_0)) - (M(f) - M(f_0))\big) \geq 2^{2j-3} r^2 t_n^2\Big) \\
&\leq \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \Big|\frac{1}{n}\sum_{i=1}^n \xi_i (f - f_0)(\mathbf{x}^{(i)})\Big| \geq 2^{2j-5} r^2 t_n^2\Big) \\
&\quad + \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \Big|\frac{1}{n}\sum_{i=1}^n (f - f_0)^2(\mathbf{x}^{(i)}) - \|f - f_0\|_{P,2}^2\Big| \geq 2^{2j-5} r^2 t_n^2\Big) \\
&\quad + \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \Big|\frac{1}{n}\sum_{i=1}^n (f - f_0)(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)}) \\
&\qquad\qquad - \mathbb{E}_{\mathbf{X}\sim P}\big[(f - f_0)(\mathbf{X}) \cdot (f_0 - f^*)(\mathbf{X})\big]\Big| \geq 2^{2j-6} r^2 t_n^2\Big) \\
&\leq \mathbb{P}\Big(\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big| \geq 2^{2j-5} r^2 \sqrt{n} t_n^2\Big) \\
&\quad + \mathbb{P}\Big(\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{P,2}^2\big)\Big| \geq 2^{2j-5} r^2 \sqrt{n} t_n^2\Big) \qquad (C.44) \\
&\quad + \mathbb{P}\Big(\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(f(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)}) \\
&\qquad\qquad - \mathbb{E}_{\mathbf{X}\sim P}\big[f(\mathbf{X}) \cdot (f_0 - f^*)(\mathbf{X})\big]\Big)\Big| \geq 2^{2j-6} r^2 \sqrt{n} t_n^2\Big).
\end{aligned}$$

Here the second inequality is due to the fact that

$$-(M(f) - M(f_0)) \geq 2^{2j-3} r^2 t_n^2$$

for $f \in \mathcal{F}_j$, and the third inequality is from the triangle inequality.

We first bound the first term of (C.44). Fix a positive integer j. By Markov's inequality,

$$\mathbb{P}\Big(\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big| \ge 2^{2j-5} r^2\sqrt{n} t_n^2\Big)$$
$$\le \frac{1}{(2^{2j-5}r^2\sqrt{n}t_n^2)^3}\cdot \mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|^3\Big]. \qquad \text{(C.45)}$$

Also, by Lemma 21 (the moment inequality for empirical processes), we have

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|^3\Big] \le C\Big(\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big]\Big)^3$$
$$+ C2^{3j}r^3t_n^3\|\xi_1\|_2^3 + Cn^{-3/2}\cdot\mathbb{E}\Big[\max_i\Big(|\xi_i|^3\cdot \sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le 2^j r t_n}} |f(\mathbf{x}^{(i)})|^3\Big)\Big]$$
$$\le C2^{3j}r^3n^{3/2}t_n^6 + C2^{3j}r^3t_n^3\|\xi_1\|_2^3 + CM^3n^{-3/2}\cdot\mathbb{E}\big[\max_i|\xi_i|^3\big].$$

Since ξ_i have finite L^q norm,

$$\mathbb{E}\big[\max_i|\xi_i|^3\big] = \mathbb{E}\big[\big(\max_i|\xi_i|^q\big)^{3/q}\big] \le \Big(\mathbb{E}\big[\max_i|\xi_i|^q\big]\Big)^{3/q} \le \Big(\sum_{i=1}^n\mathbb{E}\big[|\xi_i|^q\big]\Big)^{3/q} = n^{3/q}\|\xi_1\|_q^3,$$

where the first inequality is from Jensen's inequality. Hence, we have

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|^3\Big] \le C2^{3j}r^3n^{3/2}t_n^6 + C2^{3j}r^3t_n^3\|\xi_1\|_2^3 + CM^3n^{3(-1/2+1/q)}\|\xi_1\|_q^3,$$

which together with (C.45) implies that

$$\mathbb{P}\Big(\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big| \ge 2^{2j-5} r^2\sqrt{n} t_n^2\Big)$$
$$\le C\cdot\frac{1}{2^{3j}r^3} + C\cdot\frac{\|\xi_1\|_2^3}{2^{3j}r^3n^{3/2}t_n^3} + C\cdot\frac{M^3\|\xi_1\|_q^3}{2^{6j}r^6n^{3(1-1/q)}t_n^6}.$$

Now, we bound the second term of (C.44). We start by applying Markov's inequality as in the case of the first term:

$$\mathbb{P}\Big(\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{P,2}^2\big)\Big| \ge 2^{2j-5} r^2\sqrt{n} t_n^2\Big)$$

$$\leq \frac{1}{(2^{2j-5}r^2\sqrt{n}t_n^2)^3} \cdot \mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{P,2}^2 \big) \Big|^3 \Big]. \tag{C.46}$$

Using the standard argument of symmetrization (see, e.g., [150, Lemma 2.3.1] or [59, Theorem 16.1]), we can show that

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{P,2}^2 \big) \Big|^3 \Big] \leq 8\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)}) \Big|^3 \Big].$$

Also, by Lemma 21, we have

$$\begin{aligned} \mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)}) \Big|^3 \Big] &\leq C\Big(\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)}) \Big| \Big] \Big)^3 \\ &\quad + C\Big(\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \|f^2\|_{P,2} \Big)^3 + Cn^{-3/2} \cdot \mathbb{E}\Big[\max_i \sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} |f(\mathbf{x}^{(i)})|^6 \Big]. \end{aligned}$$

Due to the contraction principle (see, e.g., [150, Proposition A.3.2]),

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)}) \Big| \Big] \leq 4M \cdot \mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)}) \Big| \Big] \leq 2^{j+2} r M \sqrt{n} t_n^2.$$

Moreover, since $\|f\|_\infty \leq M$ for every $f \in \mathcal{F} - \{f_0\}$,

$$\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \|f^2\|_{P,2} \leq M \cdot \sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \|f\|_{P,2} \leq 2^j r M t_n.$$

Therefore, we have

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)}) \Big|^3 \Big] \leq C 2^{3j} r^3 M^3 n^{3/2} t_n^6 + C 2^{3j} r^3 M^3 t_n^3 + C M^6 n^{-3/2},$$

which directly leads to

$$\begin{aligned} &\mathbb{P}\Big(\sup_{\substack{f\in\mathcal{F}-\{f_0\} \\ \|f\|_{P,2}\leq 2^j r t_n}} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{P,2}^2 \big) \Big| \geq 2^{2j-5} r^2 \sqrt{n} t_n^2 \Big) \\ &\qquad \leq C \cdot \frac{M^3}{2^{3j} r^3} + C \cdot \frac{M^3}{2^{3j} r^3 n^{3/2} t_n^3} + C \cdot \frac{M^6}{2^{6j} r^6 n^3 t_n^6}, \end{aligned}$$

because of (C.46).

The third term of (C.44) can be bounded in the same way as the second term. Indeed, by repeating the same argument (except for the application of the contraction principle), we can show that

$$\begin{aligned}\mathbb{P}\Bigg(\sup_{\substack{f\in\mathcal{F}-\{f_0\}\\ \|f\|_{P,2}\le 2^j r t_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(f(\mathbf{x}^{(i)})\cdot(f_0-f^*)(\mathbf{x}^{(i)}) \\ -\mathbb{E}_{\mathbf{X}\sim P}\big[f(\mathbf{X})\cdot(f_0-f^*)(\mathbf{X})\big]\Big)\Big| \ge 2^{2j-6}r^2\sqrt{n}t_n^2\Bigg) \\ \le C\cdot\frac{1}{2^{3j}r^3}+C\cdot\frac{\|f_0-f^*\|_\infty^3}{2^{3j}r^3n^{3/2}t_n^3}+C\cdot\frac{M^3\cdot\|f_0-f^*\|_\infty^3}{2^{6j}r^6n^3t_n^6}.\end{aligned}$$

As a result, by (C.43) and (C.44),

$$\begin{aligned}\mathbb{P}(\|\hat{f}-f_0\|_{P,2}>rt_n) &\le \epsilon+\sum_{j=1}^{\infty}\Bigg[C\cdot\frac{1+M^3}{2^{3j}r^3}+C\cdot\frac{\|\xi_1\|_2^3+\|f_0-f^*\|_\infty^3+M^3}{2^{3j}r^3n^{3/2}t_n^3} \\ &\qquad + C\cdot\frac{M^3(\|\xi_1\|_q^3n^{3/q}+\|f_0-f^*\|_\infty^3+M^3)}{2^{6j}r^6n^3t_n^6}\Bigg] \\ &\le \epsilon+C\cdot\frac{1+M^3}{r^3}+C\cdot\frac{\|\xi_1\|_2^3+\|f_0-f^*\|_\infty^3+M^3}{r^3n^{3/2}t_n^3} \\ &\qquad + C\cdot\frac{M^3(\|\xi_1\|_q^3n^{3/q}+\|f_0-f^*\|_\infty^3+M^3)}{r^6n^3t_n^6},\end{aligned}$$

and replacing r with t/t_n completes the proof. □

C.19 Proof of Lemma 15

Proof of Lemma 15. We first observe that every function $f\in\mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ can be uniquely split into a multi-affine term and the remaining term as follows. Let $\mathcal{A}$ be the collection of all multi-affine functions

$$a(x_1,\ldots,x_d)=\sum_{0\le|S|\le s}a_S\prod_{j\in S}x_j,$$

where $a_S\in\mathbb{R}$ for each $S\subseteq[d]$ with $0\le|S|\le s$, and let $\mathcal{H}$ be the collection of all functions

$$h(x_1,\ldots,x_d)=\sum_{0<|S|\le s}\int_{[0,1)^{|S|}\setminus\{\mathbf{0}\}}\prod_{j\in S}(x_j-t_j)_+\,d\nu_S(t_j,j\in S),$$

where ν_S is a finite signed measure on $[0,1)^{|S|}\setminus\{\mathbf{0}\}$ for each $S\subseteq[d]$ with $0<|S|\le s$. By the uniqueness of representation (3.3) (see Lemma 8), clearly, we have

$$\mathcal{F}^{d,s}_{\infty-\mathrm{mars}}=\mathcal{A}\oplus\mathcal{H},$$

where $\oplus$ indicates direct sum. Also, note that if $f = a + h$ where $a \in \mathcal{A}$ and $h \in \mathcal{H}$ are in the forms above, then

$$\|h\|_\infty \leq \sum_{0<|S|\leq s} |\nu_S|\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big) = V_{\mathrm{mars}}(f).$$

Write $\hat{f}^{d,s}_{n,V} = \hat{a}^{d,s}_{n,V} + \hat{h}^{d,s}_{n,V}$ and $f^* = a^* + h^*$ where $\hat{a}^{d,s}_{n,V}, a^* \in \mathcal{A}$ and $\hat{h}^{d,s}_{n,V}, h^* \in \mathcal{H}$. Because

$$\|\hat{h}^{d,s}_{n,V}\|_\infty \leq V_{\mathrm{mars}}\big(\hat{f}^{d,s}_{n,V}\big) = V \quad \text{and} \quad \|h^*\|_\infty \leq V_{\mathrm{mars}}(f^*) = V,$$

it follows that

$$\|\hat{f}^{d,s}_{n,V} - f^*\|_\infty \leq \|\hat{a}^{d,s}_{n,V} - a^*\|_\infty + \|\hat{h}^{d,s}_{n,V}\|_\infty + \|h^*\|_\infty \leq \|\hat{a}^{d,s}_{n,V} - a^*\|_\infty + 2V.$$

Hence, in order to prove the lemma, it suffices to show that

$$\|\hat{a}^{d,s}_{n,V} - a^*\|_\infty = O_p(1).$$

Recall from (3.5) that there is no constraint on multi-affine terms in our definition of $\hat{f}^{d,s}_{n,V}$. Thus, we can characterize $\hat{a}^{d,s}_{n,V}$ as

$$\begin{aligned}
\hat{a}^{d,s}_{n,V} &= \operatorname*{argmin}_{a\in\mathcal{A}} \left\{ \sum_{i=1}^n \big(y_i - \hat{h}^{d,s}_{n,V}(\mathbf{x}^{(i)}) - a(\mathbf{x}^{(i)})\big)^2 \right\} \\
&= \operatorname*{argmin}_{a\in\mathcal{A}} \left\{ \sum_{i=1}^n \big(\xi_i - (\hat{h}^{d,s}_{n,V} - h^*)(\mathbf{x}^{(i)}) - (a - a^*)(\mathbf{x}^{(i)})\big)^2 \right\}. \qquad \text{(C.47)}
\end{aligned}$$

Let $\mathbf{X}$ be the matrix whose rows and columns are indexed by $i \in \{1, \dots, n\}$ and $S \subseteq [d]$ with $0 \leq |S| \leq s$, where $X_{i\emptyset} = 1$ and

$$X_{iS} = \prod_{j\in S} x^{(i)}_j$$

for $S \neq \emptyset$. Also, consider the coefficient vectors $\hat{\mathbf{a}}$ and $\mathbf{a}^*$ of the multi-affine functions $\hat{a}^{d,s}_{n,V}$ and a^*. Specifically,

$$\hat{\mathbf{a}} = (\hat{a}_S, S \subseteq [d], 0 \leq |S| \leq s) \quad \text{and} \quad \mathbf{a}^* = (a^*_S, S \subseteq [d], 0 \leq |S| \leq s),$$

where

$$\hat{a}^{d,s}_{n,V}(x_1, \dots, x_d) = \sum_{0\leq|S|\leq s} \hat{a}_S \prod_{j\in S} x_j \quad \text{and} \quad a^*(x_1, \dots, x_d) = \sum_{0\leq|S|\leq s} a^*_S \prod_{j\in S} x_j.$$

Moreover, let $\hat{\mathbf{h}}$ and $\mathbf{h}^*$ be the n-dimensional vectors for which

$$\hat{h}_i = \hat{h}^{d,s}_{n,V}(\mathbf{x}^{(i)}) \quad \text{and} \quad h^*_i = h^*(\mathbf{x}^{(i)})$$

for $i = 1, \ldots, n$. Then, it follows from (C.47) that

$$\hat{\mathbf{a}} = \operatorname*{argmin}_{\mathbf{a}} \big\{ \|\boldsymbol{\xi} - (\hat{\mathbf{h}} - \mathbf{h}^*) - \mathbf{X}(\mathbf{a} - \mathbf{a}^*)\|_2^2 : \mathbf{a} = (a_S, S \subseteq [d], 0 \leq |S| \leq s) \big\}, \tag{C.48}$$

where $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_n)$.

Now, let $\Pi_{\mathbf{X}}$ be the projection operator onto the column space of $\mathbf{X}$. Then, (C.48) gives

$$\mathbf{X}(\hat{\mathbf{a}} - \mathbf{a}^*) = \Pi_{\mathbf{X}}\big(\boldsymbol{\xi} - (\hat{\mathbf{h}} - \mathbf{h}^*)\big).$$

Since $\Pi_{\mathbf{X}}$ is a projection operator,

$$\begin{aligned} \big\|\Pi_{\mathbf{X}}\big(\boldsymbol{\xi} - (\hat{\mathbf{h}} - \mathbf{h}^*)\big)\big\|_2 \leq \|\boldsymbol{\xi} - (\hat{\mathbf{h}} - \mathbf{h}^*)\|_2 &\leq \left(\sum_{i=1}^n \xi_i^2\right)^{1/2} + \sqrt{n}\big(\|\hat{h}^{d,s}_{n,V}\|_\infty + \|h^*\|_\infty\big) \\ &\leq \left(\sum_{i=1}^n \xi_i^2\right)^{1/2} + 2V \cdot \sqrt{n}. \end{aligned}$$

Also, we have that

$$\|\mathbf{X}(\hat{\mathbf{a}} - \mathbf{a}^*)\|_2^2 = (\hat{\mathbf{a}} - \mathbf{a}^*)^T \mathbf{X}^T \mathbf{X} (\hat{\mathbf{a}} - \mathbf{a}^*) \geq \lambda_{\min}(\mathbf{X}^T\mathbf{X}) \cdot \|\hat{\mathbf{a}} - \mathbf{a}^*\|_2^2,$$

where $\lambda_{\min}(\mathbf{X}^T\mathbf{X})$ is the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}$. Hence, combining these results, we can derive that

$$\big(\lambda_{\min}(\mathbf{X}^T\mathbf{X}/n)\big)^{1/2} \cdot \|\hat{\mathbf{a}} - \mathbf{a}^*\|_2 \leq \left(\frac{1}{n}\sum_{i=1}^n \xi_i^2\right)^{1/2} + 2V. \tag{C.49}$$

Consider the symmetric matrix $\boldsymbol{\Sigma}$ whose rows and columns are indexed by $S \subseteq [d]$ with $0 \leq |S| \leq s$, where

$$\Sigma_{SS'} = \mathbb{E}\bigg[\prod_{j \in S} x_j^{(1)} \cdot \prod_{j' \in S'} x_{j'}^{(1)}\bigg].$$

By the law of large numbers, for every $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\big(\|\mathbf{X}^T\mathbf{X}/n - \boldsymbol{\Sigma}\|_F > \epsilon\big) = 0,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Also,

$$\big|\lambda_{\min}(\mathbf{X}^T\mathbf{X}/n) - \lambda_{\min}(\boldsymbol{\Sigma})\big| \leq \|\mathbf{X}^T\mathbf{X}/n - \boldsymbol{\Sigma}\|_F$$

due to Weyl's inequality, and the following lemma, which we prove in Appendix C.24, ensures that $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$.

Lemma 22. *The smallest eigenvalue of $\boldsymbol{\Sigma}$ is positive, i.e., $\lambda_{min}(\boldsymbol{\Sigma}) > 0$.*

For these reasons,

$$\mathbb{P}\Big(\lambda_{\min}(\mathbf{X}^T\mathbf{X}/n) < \frac{\lambda_{\min}(\mathbf{\Sigma})}{2}\Big) \le \mathbb{P}\Big(\big|\lambda_{\min}(\mathbf{X}^T\mathbf{X}/n) - \lambda_{\min}(\mathbf{\Sigma})\big| > \frac{\lambda_{\min}(\mathbf{\Sigma})}{2}\Big)$$
$$\le \mathbb{P}\Big(\|\mathbf{X}^T\mathbf{X}/n - \mathbf{\Sigma}\|_F > \frac{\lambda_{\min}(\mathbf{\Sigma})}{2}\Big),$$

which leads to

$$\lim_{n\to\infty} \mathbb{P}\Big(\lambda_{\min}(\mathbf{X}^T\mathbf{X}/n) < \frac{\lambda_{\min}(\mathbf{\Sigma})}{2}\Big) = 0.$$

Consequently, by (C.49), we have

$$\limsup_{n\to\infty} \mathbb{P}\big(\|\hat{\mathbf{a}} - \mathbf{a}^*\|_2 > K\big) \le \lim_{n\to\infty} \mathbb{P}\Big(\lambda_{\min}(\mathbf{X}^T\mathbf{X}/n) < \frac{\lambda_{\min}(\mathbf{\Sigma})}{2}\Big)$$
$$+ \limsup_{n\to\infty} \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i^2 > \Big\{\Big(\frac{\lambda_{\min}(\mathbf{\Sigma})}{2}\Big)^{1/2} \cdot K - 2V\Big\}^2\Big)$$
$$\le \Big\{\Big(\frac{\lambda_{\min}(\mathbf{\Sigma})}{2}\Big)^{1/2} \cdot K - 2V\Big\}^{-2} \cdot \|\xi_1\|_2^2$$

provided that $(\lambda_{\min}(\mathbf{\Sigma})/2)^{1/2} \cdot K - 2V > 0$. Here, Markov's inequality is used for the second inequality. This result implies that for each $\epsilon > 0$, there exists $K > 0$ such that

$$\mathbb{P}\big(\|\hat{\mathbf{a}} - \mathbf{a}^*\|_2 > K\big) < \epsilon,$$

i.e., $\|\hat{\mathbf{a}} - \mathbf{a}^*\|_2 = O_p(1)$. By Cauchy inequality,

$$\|\hat{a}_{n,V}^{d,s} - a^*\|_\infty \le \|\hat{\mathbf{a}} - \mathbf{a}^*\|_1 \le C_d \cdot \|\hat{\mathbf{a}} - \mathbf{a}^*\|_2,$$

and thus, it follows that $\|\hat{a}_{n,V}^{d,s} - a^*\|_\infty = O_p(1)$. □

C.20 Proof of Lemma 16

Our proof of Lemma 16 is based on the following maximal inequality for empirical processes involving the Bernstein norm. For a random variable X with law P on $\mathcal{X}$ and a real-valued function f on $\mathcal{X}$, the Bernstein norm $\|f\|_{P,B}$ of f is defined as

$$\|f\|_{P,B} = \big(2\mathbb{E}_P\big[\exp(|f(\mathbf{X})|) - 1 - |f(\mathbf{X})|\big]\big)^{1/2}.$$

In fact, the Bernstein norm is not a norm because it is not homogeneous and does not satisfy the triangle inequality. However, we can still use it for measuring the size of functions.

Lemma 23 (Lemma 3.4.3 of [150]). *Suppose* $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$ *are i.i.d. random variables with law* P *on* $\mathcal{X}$*, and* $\mathcal{F}$ *is a countable collection of real-valued functions on* $\mathcal{X}$*. Also, assume that* $\|f\|_{P,B} \leq \delta$ *for every* $f \in \mathcal{F}$*. Then, we have*

$$\mathbb{E}_P\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} f(\mathbf{x}^{(i)})\Big|\Big] \leq C J_{[\,]}(\delta, \mathcal{F}, \|\cdot\|_{P,B})\bigg(1 + \frac{J_{[\,]}(\delta, \mathcal{F}, \|\cdot\|_{P,B})}{\delta^2\sqrt{k}}\bigg)$$

for some universal positive constant C*.*

Remark 14. *As in Theorem 24, the countability assumption on* $\mathcal{F}$ *in Lemma 23 is just for the measurability of the supremum. This condition can be dropped if* $\mathcal{F}$ *is pointwise measurable; that is,* $\mathcal{F}$ *has a countable subset* $\mathcal{G}$ *for which for every* $f \in \mathcal{F}$*, there exists a sequence* $\{g_m\}_{m\geq 1}$ *in* $\mathcal{G}$ *such that* $g_m(\mathbf{x}) \to f(\mathbf{x})$ *for every* $\mathbf{x} \in \mathcal{X}$*. In this case,*

$$\sup_{f\in\mathcal{F}}\Big|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} f(\mathbf{x}^{(i)})\Big| = \sup_{g\in\mathcal{G}}\Big|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} g(\mathbf{x}^{(i)})\Big|,$$

so that the measurability issue can be avoided, and the lemma is still valid without the assumption of countability.

Proof of Lemma 16. Let P be the law of $(\mathbf{x}^{(i)}, \epsilon_i)$ on $[0,1]^d \times \mathbb{R}$. Also, let $\mathcal{A}_t$ be the collection of all functions on $[0,1]^d \times \mathbb{R}$ of the form

$$\Phi_f(\mathbf{X}, \epsilon) = \frac{1}{2M} \cdot \epsilon f(\mathbf{X}) \quad \text{for } \mathbf{X} \in [0,1]^d \text{ and } \epsilon \in \mathbb{R},$$

where $f \in \mathcal{F}(V, M) - \{f^*\}$ and $\|f\|_{p_0,2} \leq t$, i.e., $f \in B_{\mathcal{F}(V,M)-\{f^*\}}(t, 0, \|\cdot\|_{p_0,2}) := B(t, 0, \|\cdot\|_{p_0,2})$. Since $B(t, 0, \|\cdot\|_{p_0,2})$ is a subset of $C([0,1]^d)$, and $(C([0,1]^d), \|\cdot\|_\infty)$ is a separable metric space, $B(t, 0, \|\cdot\|_{p_0,2})$ is also separable with respect to the sup-norm. Let $\mathcal{G}$ be a countable dense subset of $B(t, 0, \|\cdot\|_{p_0,2})$ with respect to the sup-norm. For every $\Phi_f \in \mathcal{A}_t$, if we let $\{g_m\}_{m\geq 1}$ be a sequence in $\mathcal{G}$ for which $\|g_m - f\|_\infty \to 0$, then we have

$$\Phi_{g_m}(\mathbf{X}, \epsilon) - \Phi_f(\mathbf{X}, \epsilon) = \frac{1}{2M} \cdot \epsilon(g_m - f)(\mathbf{X}) \to 0 \quad \text{for every } (\mathbf{X}, \epsilon) \in [0,1]^d \times \mathbb{R}.$$

Hence, $\mathcal{A}_t$ is pointwise measurable, and Lemma 23 is valid for $\mathcal{A}_t$.

Note that for $\Phi_f \in \mathcal{A}_t$,

$$\begin{aligned} \|\Phi_f\|_{P,B} &= \bigg(2\mathbb{E}_P\Big[\exp\Big(\Big|\frac{1}{2M} \cdot \epsilon f(\mathbf{X})\Big|\Big) - 1 - \Big|\frac{1}{2M} \cdot \epsilon f(\mathbf{X})\Big|\Big]\bigg)^{1/2} \qquad \text{(C.50)} \\ &= \bigg(2\sum_{m=2}^{\infty}\frac{1}{m!} \cdot \mathbb{E}_P\Big[\Big(\frac{|f(\mathbf{X})|}{2M}\Big)^m\Big]\bigg)^{1/2} \leq \bigg(2\sum_{m=2}^{\infty}\frac{1}{m!} \cdot \mathbb{E}_P\Big[\Big(\frac{|f(\mathbf{X})|}{2M}\Big)^2\Big]\bigg)^{1/2} \leq \frac{c_0 t}{2M}, \end{aligned}$$

where $c_0 := (2(e-2))^{1/2}$. Here, the first inequality is due to the fact that $\|f\|_\infty \le M$ for every $f \in \mathcal{F}(V,M) - \{f^*\}$, and the second inequality is from the fact that $\|f\|_{p_0,2} \le t$. Thus, by Lemma 23, we have

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le t}} \Big|\frac{1}{\sqrt{k}}\sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] = 2M\cdot\mathbb{E}\Big[\sup_{\Phi_f\in\mathcal{A}_t}\Big|\frac{1}{\sqrt{k}}\sum_{i=1}^k \Phi_f(\mathbf{x}^{(i)},\epsilon_i)\Big|\Big]$$
$$\le C\cdot 2M\cdot J_{[\,]}\Big(\frac{c_0 t}{2M},\mathcal{A}_t,\|\cdot\|_{P,B}\Big)\bigg(1+\frac{J_{[\,]}(c_0t/(2M),\mathcal{A}_t,\|\cdot\|_{P,B})}{(c_0t/(2M))^2\sqrt{k}}\bigg). \tag{C.51}$$

Now, we relate the bracketing entropy of $\mathcal{A}_t$ with respect to the Bernstein norm to that of $\mathcal{F}(V,M)-\{f^*\}$ with respect to the norm $\|\cdot\|_{p_0,2}$. Suppose $\Phi_f\in\mathcal{A}_t$ and let $[f_1,f_2]$ be a bracket containing f. Since $\|f\|_\infty\le M$, we can assume that $\|f_1\|_\infty\le M$ and $\|f_2\|_\infty\le M$. Also, let p and q denote the functions

$$p(\epsilon)=\epsilon_+=\max\{\epsilon,0\}\quad\text{and}\quad q(\epsilon)=\epsilon_-=\max\{-\epsilon,0\}.$$

Then, clearly,

$$\Phi_f\in\Big[\frac{1}{2M}\{f_1\otimes p-f_2\otimes q\},\frac{1}{2M}\{f_2\otimes p-f_1\otimes q\}\Big].$$

Recall that $\otimes$ denotes the tensor product between functions. For example, $f_1\otimes p$ indicates the function defined by

$$(f_1\otimes p)(\mathbf{X},\epsilon)=f_1(\mathbf{X})\cdot p(\epsilon)$$

for $\mathbf{X}\in[0,1]^d$ and $\epsilon\in\mathbb{R}$. Also, observe that by the same argument as in (C.50),

$$\Big\|\frac{1}{2M}\{f_2\otimes p-f_1\otimes q\}-\frac{1}{2M}\{f_1\otimes p-f_2\otimes q\}\Big\|_{P,B}=\Big\|\frac{1}{2M}(f_2-f_1)\otimes(p+q)\Big\|_{P,B}$$
$$=\bigg(2\mathbb{E}_P\Big[\exp\Big(\Big|\frac{1}{2M}(f_2-f_1)(\mathbf{X})\Big|\Big)-1-\Big|\frac{1}{2M}(f_2-f_1)(\mathbf{X})\Big|\Big]\bigg)^{1/2}\le\frac{c_0}{2M}\cdot\|f_2-f_1\|_{p_0,2}.$$

For these reasons, for every $\epsilon>0$, we have

$$N_{[\,]}\Big(\frac{c_0\epsilon}{2M},\mathcal{A}_t,\|\cdot\|_{P,B}\Big)\le N_{[\,]}(\epsilon,\mathcal{F}(V,M)-\{f^*\},\|\cdot\|_{p_0,2}),$$

which implies that

$$\begin{aligned}J_{[\,]}\Big(\frac{c_0t}{2M},\mathcal{A}_t,\|\cdot\|_{P,B}\Big)&=\int_0^{c_0t/(2M)}\sqrt{1+N_{[\,]}(\epsilon,\mathcal{A}_t,\|\cdot\|_{P,B})}\,d\epsilon\\&=\frac{c_0}{2M}\int_0^t\sqrt{1+N_{[\,]}\Big(\frac{c_0\epsilon}{2M},\mathcal{A}_t,\|\cdot\|_{P,B}\Big)}\,d\epsilon\\&\le\frac{c_0}{2M}\int_0^t\sqrt{1+N_{[\,]}(\epsilon,\mathcal{F}(V,M)-\{f^*\},\|\cdot\|_{p_0,2})}\,d\epsilon\end{aligned}$$

$$= \frac{c_0}{2M} \cdot J_{[\,]}(t, \mathcal{F}(V, M) - \{f^*\}, \| \cdot \|_{p_0,2}).$$

As a result, we can deduce from (C.51) that

$$\mathbb{E}\Bigg[\sup_{\substack{f \in \mathcal{F}(V,M) - \{f^*\} \\ \|f\|_{p_0,2} \le t}} \Big| \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \epsilon_i f(\mathbf{x}^{(i)}) \Big| \Bigg]$$
$$\le C \cdot J_{[\,]}(t, \mathcal{F}(V, M) - \{f^*\}, \| \cdot \|_{p_0,2}) \cdot \left(1 + M \cdot \frac{J_{[\,]}(t, \mathcal{F}(V, M) - \{f^*\}, \| \cdot \|_{p_0,2})}{t^2 \sqrt{k}} \right).$$

□

C.21 Proof of Lemma 17

Proof of Lemma 17. Suppose $f = f_{c,\{\nu_S\}} \in \mathcal{F}(V, M) - \{f^*\}$. By the definition of $\mathcal{F}(V, M)$, we have $\|f\|_\infty \le M$ and $V_{\mathrm{mars}}(f) \le V + V_{\mathrm{mars}}(f^*) \le 2V$. Let $c_\emptyset = c$, and let $c_S = \nu_S(\{\mathbf{0}\})$ for each $S \subseteq [d]$ with $0 < |S| \le s$. Then, the function f can be written as

$$f(x_1, \dots, x_d) = c + \sum_{0 < |S| \le s} \int_{[0,1)^{|S|}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S)$$
$$= \sum_{0 \le |S| \le s} c_S \prod_{j \in S} x_j + \sum_{0 < |S| \le s} \int_{[0,1)^{|S|} \setminus \{\mathbf{0}\}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S).$$

As in the proof of Lemma 12, we first bound c_S for each $S \subseteq [d]$ with $0 \le |S| \le s$.

First, note that $|c_\emptyset| = |f(\mathbf{0})| \le M$. Also, since

$$f(1, 0, \dots, 0) = c_\emptyset + c_{\{1\}} + \int_{(0,1)} (1 - t_1)_+ \, d\nu_{\{1\}}(t_1),$$

we have

$$|c_{\{1\}}| \le |f(1, 0, \dots, 0)| + |c_\emptyset| + \Big| \int_{(0,1)} (1 - t_1)_+ \, d\nu_{\{1\}}(t_1) \Big|$$
$$\le |f(1, 0, \dots, 0)| + |c_\emptyset| + |\nu_{\{1\}}|((0, 1)) \le 2M + 2V.$$

By the same argument, we can show that

$$|c_S| \le 2M + 2V$$

for every $S \subseteq [d]$ with $|S| = 1$. Moreover, since

$$f(1, 1, 0, \dots, 0) = c_\emptyset + c_{\{1\}} + c_{\{2\}} + c_{\{1,2\}} + \int_{(0,1)} (1 - t_1)_+ \, d\nu_{\{1\}}(t_1) + \int_{(0,1)} (1 - t_2)_+ \, d\nu_{\{2\}}(t_2)$$

$$+ \int_{[0,1)^2 \setminus \{\mathbf{0}\}} (1-t_1)_+ (1-t_2)_+ \, d\nu_{\{1,2\}}(t_1, t_2),$$

it follows that

$$\begin{aligned}
|c_{\{1,2\}}| &\leq |f(1,1,0,\ldots,0)| + |c_\emptyset| + |c_{\{1\}}| + |c_{\{2\}}| \\
&\quad + \Big| \int_{(0,1)} (1-t_1)_+ \, d\nu_{\{1\}}(t_1) \Big| + \Big| \int_{(0,1)} (1-t_2)_+ \, d\nu_{\{2\}}(t_2) \Big| \\
&\quad + \Big| \int_{[0,1)^2 \setminus \{\mathbf{0}\}} (1-t_1)_+ (1-t_2)_+ \, d\nu_{\{1,2\}}(t_1, t_2) \Big| \\
&\leq |f(1,1,0,\ldots,0)| + |c_\emptyset| + |c_{\{1\}}| + |c_{\{2\}}| \\
&\quad + |\nu_{\{1\}}|((0,1)) + |\nu_{\{2\}}|((0,1)) + |\nu_{\{1,2\}}|\big([0,1)^2 \setminus \{\mathbf{0}\}\big) \\
&\leq M + M + (2M + 2V) + (2M + 2V) + 2V = 6M + 6V.
\end{aligned}$$

Similarly, it can be shown that

$$|c_S| \leq 6M + 6V$$

for all $S \subseteq [d]$ with $|S| = 2$. Repeating this argument, we can show inductively that

$$|c_S| \leq C_s(M + V) =: \widetilde{M}$$

for every $S \subseteq [d]$ with $0 \leq |S| \leq s$.

Now, we repeat our arguments in the proof of Lemma 12. We cover $\mathcal{F}(V, M) - \{f^*\}$ with function classes whose upper bounds of bracketing entropy can be obtained from those of $\mathcal{D}_m$ for $m \geq 1$. Fix $\delta > 0$, which we will specify later, and let

$$\mathcal{K} = \big\{(k_S, S \subseteq [d], 0 \leq |S| \leq s) : -(K+1) \leq k_S \leq K \text{ for all } S \subseteq [d] \text{ with } 0 \leq |S| \leq s\big\}$$

where $K = \lfloor \widetilde{M}/\delta \rfloor$. For each $\mathbf{k} \in \mathcal{K}$, we also let

$$\mathcal{G}_{\mathbf{k}} = \big\{f_{c, \{\nu_S\}} \in \mathcal{F}(V, M) - \{f^*\} : k_S \delta \leq c_S \leq (k_S + 1)\delta \text{ for each } S \subseteq [d] \text{ with } 0 \leq |S| \leq s\big\}.$$

Then, by definition,

$$\mathcal{F}(V, M) - \{f^*\} = \bigcup_{\mathbf{k} \in \mathcal{K}} \mathcal{G}_{\mathbf{k}},$$

from which we can obtain

$$\begin{aligned}
\log N_{[\,]}(\epsilon, \mathcal{F}(V, M) - \{f^*\}, \|\cdot\|_{p_0,2}) &\leq \log \Big(\sum_{\mathbf{k} \in \mathcal{K}} N_{[\,]}(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_{p_0,2}) \Big) \\
&\leq \log |\mathcal{K}| + \sup_{\mathbf{k} \in \mathcal{K}} \log N_{[\,]}(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_{p_0,2}) \qquad \text{(C.52)} \\
&\leq 2^d \log \Big(2 + \frac{2\widetilde{M}}{\delta} \Big) + \sup_{\mathbf{k} \in \mathcal{K}} \log N_{[\,]}(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_{p_0,2}).
\end{aligned}$$

Fix $\mathbf{k} \in \mathcal{K}$. Let $\mathcal{G}_{\mathbf{k}}^{\emptyset}$ be the collection of all constant functions on $[0,1]^d$

$$(x_1, \dots, x_d) \mapsto c$$

where $k_{\emptyset}\delta \leq c \leq (k_{\emptyset}+1)\delta$, and for each $S \subseteq [d]$ with $0 < |S| \leq s$, let $\mathcal{G}_{\mathbf{k}}^{S}$ be the collection of all functions on $[0,1]^d$ of the form

$$(x_1, \dots, x_d) \mapsto c_S \prod_{j \in S} x_j + \int_{[0,1)^{|S|} \setminus \{\mathbf{0}\}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S),$$

where $k_S\delta \leq c_S \leq (k_S+1)\delta$ and ν_S is a signed measure on $[0,1)^{|S|}$ with $|\nu_S|([0,1)^{|S|} \setminus \{\mathbf{0}\}) \leq 2V$. It then follows that

$$\mathcal{G}_{\mathbf{k}} \subseteq +_{0 \leq |S| \leq s} \mathcal{G}_{\mathbf{k}}^{S},$$

which leads to

$$\log N_{[\,]}(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_{p_0,2}) \leq \sum_{0 \leq |S| \leq s} \log N_{[\,]}\Big(\frac{\epsilon}{2^d}, \mathcal{G}_{\mathbf{k}}^{S}, \|\cdot\|_{p_0,2}\Big). \tag{C.53}$$

It is clear that

$$N_{[\,]}(\epsilon, \mathcal{G}_{\mathbf{k}}^{\emptyset}, \|\cdot\|_{p_0,2}) \leq 1 + \frac{\delta}{\epsilon}. \tag{C.54}$$

It is thus enough to bound $\log N_{[\,]}(\epsilon, \mathcal{G}_{\mathbf{k}}^{S}, \|\cdot\|_{p_0,2})$ for each $S \subseteq [d]$ with $0 < |S| \leq s$. For each $S \subseteq [d]$ with $0 < |S| \leq s$, let $\widetilde{\mathcal{G}}_S$ be the collection of all functions on $[0,1]^{|S|}$ of the form

$$(x_j, j \in S) \mapsto \int_{[0,1)^{|S|}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S)$$

where ν_S is a signed measure on $[0,1)^{|S|}$ with total variation $|\nu_S|([0,1)^{|S|}) \leq 2V + \delta$. As in the proof of Lemma 12, we can easily show that

$$N_{[\,]}(\epsilon, \mathcal{G}_{\mathbf{k}}^{S}, \|\cdot\|_{p_0,2}) \leq N_{[\,]}(\epsilon, \widetilde{\mathcal{G}}_S, \|\cdot\|_{\widetilde{p}_0,2}), \tag{C.55}$$

where $\widetilde{p}_0$ is the probability density function on $[0,1]^{|S|}$ defined by

$$\widetilde{p}_0(x_j, j \in S) = \int_{[0,1]^{d-|S|}} p_0(x_1, \dots, x_d) \, d(x_j, j \in [d] \setminus S)$$

for $(x_j, j \in S) \in [0,1]^{|S|}$. Since p_0 is bounded by B, $\widetilde{p}_0$ is also bounded by B, i.e., $\|\widetilde{p}_0\|_\infty \leq B$. Hence, from Theorem 6, we can deduce that

$$\begin{aligned}\log N_{[\,]}(\epsilon, \widetilde{\mathcal{G}}_S, \|\cdot\|_{\widetilde{p}_0,2}) &\leq \log N_{[\,]}\Big(\frac{\epsilon}{\sqrt{B}}, \widetilde{\mathcal{G}}_S, \|\cdot\|_2\Big) \\ &\leq C_{|S|}\Big(\frac{4\sqrt{B}(2V+\delta)}{\epsilon}\Big)^{1/2} \Big|\log\Big(\frac{4\sqrt{B}(2V+\delta)}{\epsilon}\Big)\Big|^{2(|S|-1)}.\end{aligned} \tag{C.56}$$

Combining (C.53), (C.54), (C.55), and (C.56), we can derive

$$\log N_{[\,]}(\epsilon, \mathcal{G}_{\mathbf{k}}, \|\cdot\|_{p_0,2}) \le \log\Big(1 + \frac{2^d\delta}{\epsilon}\Big) + C_s \sum_{0<|S|\le s} \Big(\frac{2^{d+2}\sqrt{B}(2V+\delta)}{\epsilon}\Big)^{1/2} \cdot \Big[\log\Big(\frac{2^{d+2}\sqrt{B}(2V+\delta)}{\epsilon}\Big)\Big]^{2(|S|-1)}.$$

With the choice of $\delta = \epsilon$, this together with (C.52) implies that

$$\log N_{[\,]}(\epsilon, \mathcal{F}(V,M) - \{f^*\}, \|\cdot\|_{p_0,2}) \le 2^d \log\Big(2 + C_s \cdot \frac{M+V}{\epsilon}\Big) + C_{B,d}\Big(2^{d+2} + \frac{2^{d+3}V}{\epsilon}\Big)^{1/2}\Big[\log\Big(2^{d+2} + \frac{2^{d+3}V}{\epsilon}\Big)\Big]^{2(s-1)}.$$

□

C.22 Proof of Lemma 19

Proof of (C.28). For each $\boldsymbol{\eta} \in \{-1,1\}^q$, we have

$$\begin{aligned}
|\nu_{\boldsymbol{\eta}}|\big((0,1)^s\big) &= \frac{V}{\sqrt{|P_l|}} \int_{(0,1)^s} \Big| \sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \Big(\prod_{j=1}^s \psi_{p_j,i_j}(t_j)\Big)\Big| \, d\mathbf{t} \\
&\le \frac{V}{\sqrt{|P_l|}} \bigg(\int_{(0,1)^s} \Big(\sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^s \psi_{p_j,i_j}(t_j)\Big)^2 d\mathbf{t}\bigg)^{1/2} \\
&= \frac{V}{\sqrt{|P_l|}} \bigg(\int_{(0,1)^s} \sum_{\mathbf{p}\in P_l} \Big(\sum_{\mathbf{i}\in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^s \psi_{p_j,i_j}(t_j)\Big)^2 d\mathbf{t}\bigg)^{1/2} \\
&= \frac{V}{\sqrt{|P_l|}} \bigg(\sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \int_{(0,1)^s} \Big(\prod_{j=1}^s \psi_{p_j,i_j}(t_j)\Big)^2 d\mathbf{t}\bigg)^{1/2} \\
&= \frac{V}{\sqrt{|P_l|}} \bigg(\sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \prod_{j=1}^s \int_0^1 (\psi_{p_j,i_j}(t_j))^2 \, dt_j\bigg)^{1/2} = \frac{V}{\sqrt{|P_l|}} \bigg(\sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \prod_{j=1}^s 2^{-p_j}\bigg)^{1/2} = V.
\end{aligned}$$

The inequality follows from Cauchy inequality, and the second equality is due to

$$\int_0^1 \psi_{m,k}(x)\psi_{m',k'}(x)\, dx = 0$$

for distinct m and m', and the third equality is from the fact that $\psi_{m,k}\psi_{m,k'} \equiv 0$ if k and k' are different. This directly implies that $V_{\text{mars}}(f_{\boldsymbol{\eta}}) \le V$ for every $\boldsymbol{\eta} \in \{-1,1\}^q$. □

Proof of (C.29). We first represent our function $f_{\boldsymbol{\eta}}$ in a simpler form. For a positive integer m and $k \in [2^m]$, we let $\Psi_{m,k}$ denote the real-valued function on $[0,1]$ defined by

$$\Psi_{m,k}(x) = \int_0^x \int_0^u \psi_{m,k}(t)\,dt\,du.$$

Then, it can be easily checked that $\Psi_{m,k}(x) = 0$ if $x \le (k-1)2^{-m}$ or $x \ge k2^{-m}$, $\Psi_{m,k}(x + 2^{-m-1}) = -\Psi_{m,k}(x)$ for $x \in [(k-1)2^{-m}, (k-1/2)2^{-m}]$, and $|\Psi_{m,k}(x)| \le 2^{-2m-6}$ for all $x \in [0,1]$. Using these $\Psi_{m,k}$, we can represent $f_{\boldsymbol{\eta}}$ as

$$\begin{aligned}
f_{\boldsymbol{\eta}}(x_1,\dots,x_d) &= \int_{(0,1)^s} \prod_{j=1}^s (x_j - t_j)_+ \, d\nu_{\boldsymbol{\eta}}(\mathbf{t}) \\
&= \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \int_{(0,1)^s} \prod_{j=1}^s \big((x_j - t_j)_+ \cdot \psi_{p_j,i_j}(t_j)\big)\, d\mathbf{t} \\
&= \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^s \int_0^1 (x_j - t_j)_+ \cdot \psi_{p_j,i_j}(t_j)\, dt_j \\
&= \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^s \int_0^{x_j} \int_0^{u_j} \psi_{p_j,i_j}(t_j)\, dt_j\, du_j = \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^s \Psi_{p_j,i_j}(x_j)
\end{aligned}$$

for each $\boldsymbol{\eta} \in \{-1,1\}^q$.

Now, we prove (C.29) using the above representation. Suppose $H(\boldsymbol{\eta},\boldsymbol{\eta}') = 1$ for $\boldsymbol{\eta},\boldsymbol{\eta}' \in \{-1,1\}^q$. Since $H(\boldsymbol{\eta},\boldsymbol{\eta}') = 1$, there exists a unique $(\mathbf{p},\mathbf{i}) \in Q$ such that $\eta_{\mathbf{p},\mathbf{i}} \ne \eta'_{\mathbf{p},\mathbf{i}}$, and thus

$$(f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'})(x_1,\dots,x_d) = \frac{V}{\sqrt{|P_l|}} (\eta_{\mathbf{p},\mathbf{i}} - \eta'_{\mathbf{p},\mathbf{i}}) \prod_{j=1}^s \Psi_{p_j,i_j}(x_j).$$

Hence,

$$\begin{aligned}
\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2 &= \frac{4V^2}{|P_l|} \int_{[0,1]^d} \left(\prod_{j=1}^s \Psi_{p_j,i_j}(x_j)\right)^2 p_0(\mathbf{x})\, d\mathbf{x} \\
&\le \frac{4BV^2}{|P_l|} \cdot \prod_{j=1}^s \int_0^1 \Psi_{p_j,i_j}^2(x_j)\, dx_j \le \frac{4BV^2}{|P_l|} \cdot \prod_{j=1}^s 2^{-p_j} \cdot 2^{-4p_j - 12} = \frac{BV^2}{|P_l|} \cdot 2^{-5l-12s+2}.
\end{aligned}$$

□

Proof of (C.30). Fix $\boldsymbol{\eta} \ne \boldsymbol{\eta}' \in \{-1,1\}^q$. For each positive integer m and $k \in [2^m]$, let $h_{m,k}$ be the real-valued function on $[0,1]$ defined by

$$h_{m,k}(x) = \begin{cases} 2^{m/2} & \text{if } (k-1)2^{-m} < x < (k-1/2)2^{-m} \\ -2^{m/2} & \text{if } (k-1/2)2^{-m} < x < k2^{-m} \\ 0 & \text{otherwise,} \end{cases}$$

and for each $(\mathbf{p},\mathbf{i}) \in Q$, let $H_{\mathbf{p},\mathbf{i}}$ be the real-valued function on $[0,1]^s$ defined by

$$H_{\mathbf{p},\mathbf{i}}(x_1,\ldots,x_s) = \prod_{j=1}^{s} h_{p_j,i_j}(x_j).$$

One can check that $\{H_{\mathbf{p},\mathbf{i}} : (\mathbf{p},\mathbf{i}) \in Q\}$ is an orthonormal set with respect to the L^2 inner product. Also, let $g_{\boldsymbol{\eta},\boldsymbol{\eta}'}$ denote the function on $[0,1]^s$ defined by

$$g_{\boldsymbol{\eta},\boldsymbol{\eta}'}(x_1,\ldots,x_s) = \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \left(\eta_{\mathbf{p},\mathbf{i}} - \eta'_{\mathbf{p},\mathbf{i}}\right) \prod_{j=1}^{s} \Psi_{p_j,i_j}(x_j).$$

Then, by Bessel's inequality,

$$\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2 \geq b\|g_{\boldsymbol{\eta},\boldsymbol{\eta}'}\|_2^2 \geq b \sum_{\mathbf{p}'\in P_l} \sum_{\mathbf{i}'\in I_{\mathbf{p}'}} \langle g_{\boldsymbol{\eta},\boldsymbol{\eta}'}, H_{\mathbf{p}',\mathbf{i}'}\rangle^2, \tag{C.57}$$

where $\|\cdot\|_2$ and $\langle\cdot,\cdot\rangle$ denote the L^2 norm and the L^2 inner product, respectively. Note that for each $(\mathbf{p}',\mathbf{i}') \in Q$,

$$\langle g_{\boldsymbol{\eta},\boldsymbol{\eta}'}, H_{\mathbf{p}',\mathbf{i}'}\rangle = \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} \left(\eta_{\mathbf{p},\mathbf{i}} - \eta'_{\mathbf{p},\mathbf{i}}\right) \prod_{j=1}^{s} \langle \Psi_{p_j,i_j}, h_{p'_j,i'_j}\rangle. \tag{C.58}$$

Now, we claim that for $(\mathbf{p},\mathbf{i}),(\mathbf{p}',\mathbf{i}') \in Q$,

$$\prod_{j=1}^{s} \langle \Psi_{p_j,i_j}, h_{p'_j,i'_j}\rangle = \begin{cases} 2^{-5l/2-7s} & \text{if } (\mathbf{p},\mathbf{i}) = (\mathbf{p}',\mathbf{i}') \\ 0 & \text{if } (\mathbf{p},\mathbf{i}) \neq (\mathbf{p}',\mathbf{i}'). \end{cases} \tag{C.59}$$

We first assume $(\mathbf{p},\mathbf{i}) \neq (\mathbf{p}',\mathbf{i}')$. If $\mathbf{p} \neq \mathbf{p}'$, then since $\sum_{j=1}^s p_j = l = \sum_{j=1}^s p'_j$, there exists $j \in [s]$ such that $p_j > p'_j$, and thus $h_{p'_j,i'_j}$ is constant on $((i_j-1)2^{-p_j}, i_j 2^{-p_j})$. Also, recall that $\Psi_{p_j,i_j}(x) = 0$ if $x \leq (i_j-1)2^{-p_j}$ or $x \geq i_j 2^{-p_j}$ and $\Psi_{p_j,i_j}(x + 2^{-p_j-1}) = -\Psi_{p_j,i_j}(x)$ for $x \in [(i_j-1)2^{-p_j}, (i_j-1/2)2^{-p_j}]$. For these reasons, we have $\langle \Psi_{p_j,i_j}, h_{p'_j,i'_j}\rangle = 0$. Otherwise, if $\mathbf{p} = \mathbf{p}'$, then $\mathbf{i}$ and $\mathbf{i}'$ should be distinct. Let $j \in [s]$ be an index for which $i_j \neq i'_j$. Then, $\Psi_{p_j,i_j}(x) h_{p'_j,i'_j}(x) = 0$ for all $x \in [0,1]$, which directly implies that $\langle \Psi_{p_j,i_j}, h_{p'_j,i'_j}\rangle = 0$. Next, we assume $(\mathbf{p},\mathbf{i}) = (\mathbf{p}',\mathbf{i}')$. In this case, the claim follows from the fact that for each $j \in [s]$,

$$\langle \Psi_{p_j,i_j}, h_{p_j,i_j}\rangle = \int_{(i_j-1)2^{-p_j}}^{i_j 2^{-p_j}} \Psi_{p_j,i_j}(x) h_{p_j,i_j}(x)\,dx = 2^{p_j/2} \int_{(i_j-1)2^{-p_j}}^{i_j 2^{-p_j}} |\Psi_{p_j,i_j}(x)|\,dx = 2^{-5p_j/2-7}.$$

Combining (C.57), (C.58), and (C.59), we obtain

$$\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2 \geq \frac{bV^2}{|P_l|} \cdot 2^{-5l-14s} \sum_{\mathbf{p}\in P_l} \sum_{\mathbf{i}\in I_{\mathbf{p}}} (\eta_{\mathbf{p},\mathbf{i}} - \eta'_{\mathbf{p},\mathbf{i}})^2 = \frac{bV^2}{|P_l|} \cdot 2^{-5l-14s+2} \cdot H(\boldsymbol{\eta},\boldsymbol{\eta}').$$

From this result, we can derive the conclusion

$$\min_{\boldsymbol{\eta}\neq\boldsymbol{\eta}'} \frac{\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2}{H(\boldsymbol{\eta},\boldsymbol{\eta}')} \geq \frac{bV^2}{|P_l|} \cdot 2^{-5l-14s+2}.$$

□

C.23 Proof of Lemma 20

Recall that we define $\mathcal{D}_m$ as the collection of all functions F on $[0,1]^m$ of the form

$$F(x_1,\dots,x_m) = \int (x_1-t_1)_+ \cdots (x_m-t_m)_+ \, d\mu(\mathbf{t}) = \int_{\prod_{j=1}^m [0,x_j]} (x_1-t_1)\cdots(x_m-t_m)\, d\mu(\mathbf{t}),$$

where μ is a signed measure on $[0,1]^m$ with total variation $|\mu|([0,1]^m) \le 1$. Lemma 20 is then a direct consequence of Theorem 3 and the following lemma, which connects the metric entropy of $\mathcal{D}_{m,\text{disc}}(R)$ and $\mathcal{D}_m$. Here, the covering number of $\mathcal{D}_m$ in the lemma is under the L^2 norm as in Theorem 3.

Lemma 24. *There exists a positive constant $c_{\rho,m}$ depending on ρ and m such that*

$$N(\epsilon, \mathcal{D}_{m,disc}(R), \|\cdot\|_n) \le N\Big(\frac{c_{\rho,m}\epsilon}{R}, \mathcal{D}_m, \|\cdot\|_2\Big)$$

for every $R > 0$ and $\epsilon > 0$.

Proof of Lemma 24. Define a map $\Phi : \mathcal{D}_{m,\text{disc}}(R) \to \mathcal{D}_m$ as follows. Given $f \in \mathcal{D}_{m,\text{disc}}(R)$ of the form

$$f(x_1,\dots,x_m) = \int_{[0,1)^m} (x_1-t_1)_+ \cdots (x_m-t_m)_+ \, d\nu(\mathbf{t})$$

where ν is a signed measure supported on $(\prod_{j=1}^m \mathcal{V}_j) \cap [0,1)^m$ with total variation $|\nu|([0,1)^m) \le R$, we first consider the signed measure μ on $[0,1]^m$ satisfying the following:

- μ is supported on the set

$$\prod_{j=1}^m \big\{ v_{i_j}^{(j)} + \big(1 - v_{n_j-1}^{(j)}\big) : i_j \in \{0,\dots,n_j-1\} \big\}.$$

- For $(i_1,\dots,i_m) \in I_0 := \{0,\dots,n_1-1\} \times \cdots \times \{0,\dots,n_m-1\}$,

$$\mu\Big(\Big\{\Big(v_{i_j}^{(j)} + \big(1 - v_{n_j-1}^{(j)}\big), j \in [m]\Big)\Big\}\Big) = \frac{1}{R} \cdot \nu\Big(\Big\{\Big(v_{i_j}^{(j)}, j \in [m]\Big)\Big\}\Big).$$

Note that $|\mu|([0,1]^m) \le 1$ since $|\nu|([0,1)^m) \le R$. Next, let F denote the function on $[0,1]^m$ of the form

$$F(x_1,\dots,x_m) = \int_{[0,1]^m} (x_1-t_1)_+ \cdots (x_m-t_m)_+ \, d\mu(\mathbf{t}),$$

and define $\Phi(f) = F$. Clearly, for $(x_1,\dots,x_m) \in [0,1]^m$,

$$F(x_1,\dots,x_m) = 0$$

if $x_j \le 1 - v^{(j)}_{n_j-1}$ for some $j \in [m]$, and

$$F(x_1, \dots, x_m) = \frac{1}{R} \cdot f\big(x_j - \big(1 - v^{(j)}_{n_j-1}\big), j \in [m]\big)$$

otherwise. Also, note that for

$$\mathbf{x} = \big((1 - z_j) \cdot v^{(j)}_{i_j-1} + z_j \cdot v^{(j)}_{i_j} + \big(1 - v^{(j)}_{n_j-1}\big), j \in [m]\big),$$

where $\mathbf{i} = (i_1, \dots, i_m) \in \prod_{j=1}^m \{1, \dots, n_j - 1\}$ and $0 \le z_1, \dots, z_m < 1$,

$$\begin{aligned}
F(\mathbf{x}) &= \int_{\prod_{j=1}^m [0,x_j]} (x_1 - t_1) \cdots (x_m - t_m)\, d\mu(\mathbf{t}) \\
&= \frac{1}{R} \sum_{\mathbf{l} \in I_0} \prod_{j=1}^m \big[x_j - v^{(j)}_{l_j} - \big(1 - v^{(j)}_{n_j-1}\big)\big] \cdot \prod_{j=1}^m \mathbf{1}(i_j - 1 \ge l_j) \cdot \beta_{\mathbf{l}} \\
&= \frac{1}{R} \sum_{\mathbf{l} \in I_0} \prod_{j=1}^m \big[(1 - z_j)\big(v^{(j)}_{i_j-1} - v^{(j)}_{l_j}\big) + z_j\big(v^{(j)}_{i_j} - v^{(j)}_{l_j}\big)\big] \cdot \prod_{j=1}^m \mathbf{1}(i_j - 1 \ge l_j) \cdot \beta_{\mathbf{l}} \\
&= \frac{1}{R} \sum_{\mathbf{l} \in I_0} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} \prod_{j=1}^m \big[(1 - z_j)^{1-\delta_j} z_j^{\delta_j} \big(v^{(j)}_{i_j-1+\delta_j} - v^{(j)}_{l_j}\big)\big] \cdot \prod_{j=1}^m \mathbf{1}(i_j - 1 \ge l_j) \cdot \beta_{\mathbf{l}} \\
&= \frac{1}{R} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} \left[\prod_{j=1}^m \big[(1 - z_j)^{1-\delta_j} z_j^{\delta_j}\big] \cdot \sum_{\mathbf{l} \in I_0} \prod_{j=1}^m \big(v^{(j)}_{i_j-1+\delta_j} - v^{(j)}_{l_j}\big) \cdot \prod_{j=1}^m \mathbf{1}(i_j - 1 \ge l_j) \cdot \beta_{\mathbf{l}}\right] \\
&= \frac{1}{R} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} \left[\prod_{j=1}^m \big[(1 - z_j)^{1-\delta_j} z_j^{\delta_j}\big] \cdot f\big(v^{(j)}_{i_j-1+\delta_j}, j \in [m]\big)\right].
\end{aligned}$$

Here we let

$$\beta_{\mathbf{l}} = \nu\big(\big\{\big(v^{(j)}_{l_j}, j \in [m]\big)\big\}\big) \quad \text{for } \mathbf{l} = (l_1, \dots, l_m) \in I_0,$$

for notational convenience. Hence, for $f, \widetilde{f} \in \mathcal{D}_{m,\mathrm{disc}}(R)$,

$$\begin{aligned}
\big\|\Phi(f) - \Phi(\widetilde{f})\big\|_2^2 &= \int_{[0,1]^m} \big|\Phi(f)(\mathbf{t}) - \Phi(\widetilde{f})(\mathbf{t})\big|^2\, d\mathbf{t} \\
&= \sum_{\mathbf{i} \in \prod_{j=1}^m \{1,\dots,n_j-1\}} \int_{v^{(1)}_{i_1-1} + (1 - v^{(1)}_{n_1-1})}^{v^{(1)}_{i_1} + (1 - v^{(1)}_{n_1-1})} \cdots \int_{v^{(m)}_{i_m-1} + (1 - v^{(m)}_{n_m-1})}^{v^{(m)}_{i_m} + (1 - v^{(m)}_{n_m-1})} \big|\Phi(f)(\mathbf{t}) - \Phi(\widetilde{f})(\mathbf{t})\big|^2\, dt_m \cdots dt_1 \\
&= \frac{1}{R^2} \sum_{\mathbf{i} \in \prod_{j=1}^m \{1,\dots,n_j-1\}} \prod_{j=1}^m \big(v^{(j)}_{i_j} - v^{(j)}_{i_j-1}\big) \\
&\qquad \cdot \int_{[0,1]^m} \left[\sum_{\boldsymbol{\delta} \in \{0,1\}^m} \prod_{j=1}^m \big[(1 - z_j)^{1-\delta_j} z_j^{\delta_j}\big] \cdot (f - \widetilde{f})\big(v^{(j)}_{i_j-1+\delta_j}, j \in [m]\big)\right]^2 d\mathbf{z}
\end{aligned}$$

$$\geq \frac{1}{R^2} \cdot \frac{\rho^m}{n_1 \cdots n_m} \sum_{\mathbf{i} \in \prod_{j=1}^m \{1,\ldots,n_j-1\}} \int_{[0,1]^m} \Bigg[\sum_{\boldsymbol{\delta} \in \{0,1\}^m} \prod_{j=1}^m \big[(1-z_j)^{1-\delta_j} z_j^{\delta_j}\big] \cdot (f - \widetilde{f})\big(v^{(j)}_{i_j-1+\delta_j}, j \in [m]\big)\Bigg]^2 d\mathbf{z}.$$

For the last equality, we change the variables as

$$t_j = (1 - z_j) \cdot v^{(j)}_{i_j-1} + z_j \cdot v^{(j)}_{i_j} + \big(1 - v^{(j)}_{n_j-1}\big)$$

for $j \in [m]$. For a fixed constant $0 < r < 1$, by Cauchy inequality,

$$\begin{aligned}
&\int_{[0,1]^m} \Bigg[\sum_{\boldsymbol{\delta} \in \{0,1\}^m} \prod_{j=1}^m \big[(1-z_j)^{1-\delta_j} z_j^{\delta_j}\big] \cdot (f - \widetilde{f})\big(v^{(j)}_{i_j-1+\delta_j}, j \in [m]\big)\Bigg]^2 d\mathbf{z} \\
&\quad \geq \int_{[r,1]^m} \Bigg[\sum_{\boldsymbol{\delta} \in \{0,1\}^m} \prod_{j=1}^m \big[(1-z_j)^{1-\delta_j} z_j^{\delta_j}\big] \cdot (f - \widetilde{f})\big(v^{(j)}_{i_j-1+\delta_j}, j \in [m]\big)\Bigg]^2 d\mathbf{z} \\
&\quad \geq \frac{1}{2} \int_{[r,1]^m} z_1^2 \cdots z_m^2 \cdot (f - \widetilde{f})^2\big(v^{(j)}_{i_j}, j \in [m]\big)\, d\mathbf{z} \\
&\qquad - \int_{[r,1]^m} \Bigg[\sum_{\boldsymbol{\delta} \in \{0,1\}^m \setminus \{\mathbf{1}\}} \prod_{j=1}^m \big[(1-z_j)^{1-\delta_j} z_j^{\delta_j}\big] \cdot (f - \widetilde{f})\big(v^{(j)}_{i_j-1+\delta_j}, j \in [m]\big)\Bigg]^2 d\mathbf{z} \\
&\quad \geq \frac{1}{2} \int_{[r,1]^m} z_1^2 \cdots z_m^2 \cdot (f - \widetilde{f})^2\big(v^{(j)}_{i_j}, j \in [m]\big)\, d\mathbf{z} \\
&\qquad - (2^m - 1) \sum_{\boldsymbol{\delta} \in \{0,1\}^m \setminus \{\mathbf{1}\}} \int_{[r,1]^m} \prod_{j=1}^m \big[(1-z_j)^{2(1-\delta_j)} z_j^{2\delta_j}\big] \cdot (f - \widetilde{f})^2\big(v^{(j)}_{i_j-1+\delta_j}, j \in [m]\big)\, d\mathbf{z} \\
&\quad = \frac{(1-r^3)^m}{2 \cdot 3^m} (f - \widetilde{f})^2\big(v^{(j)}_{i_j}, j \in [m]\big) \\
&\qquad - \frac{2^m - 1}{3^m} \cdot \sum_{\boldsymbol{\delta} \in \{0,1\}^m \setminus \{\mathbf{1}\}} \Big[(1-r^3)^{\sum_j \delta_j} (1-r)^{3(m - \sum_j \delta_j)} \cdot (f - \widetilde{f})^2\big(v^{(j)}_{i_j-1+\delta_j}, j \in [m]\big)\Big]
\end{aligned}$$

for all $(i_1, \ldots, i_m) \in \prod_{j=1}^m \{1, \ldots, n_j - 1\}$. Thus, for $f, \widetilde{f} \in \mathcal{D}_{m,\mathrm{disc}}(R)$,

$$\begin{aligned}
\big\|\Phi(f) - \Phi(\widetilde{f})\big\|_2^2 &\geq \Bigg[\frac{(1-r^3)^m}{2 \cdot 3^m} - \frac{2^m - 1}{3^m} \cdot \sum_{\boldsymbol{\delta} \in \{0,1\}^m \setminus \{\mathbf{1}\}} \Big[(1-r^3)^{\sum_j \delta_j} (1-r)^{3(m - \sum_j \delta_j)}\Big]\Bigg] \\
&\qquad \cdot \frac{1}{R^2} \cdot \frac{\rho^m}{n_1 \cdots n_m} \sum_{\mathbf{i} \in \prod_{j=1}^m \{1,\ldots,n_j-1\}} (f - \widetilde{f})^2\big(v^{(j)}_{i_j}, j \in [m]\big) \\
&= \Bigg[\frac{(1-r^3)^m}{2 \cdot 3^m} - \frac{2^m - 1}{3^m} \cdot \sum_{\boldsymbol{\delta} \in \{0,1\}^m \setminus \{\mathbf{1}\}} \Big[(1-r^3)^{\sum_j \delta_j} (1-r)^{3(m - \sum_j \delta_j)}\Big]\Bigg] \cdot \frac{\rho^m}{R^2} \cdot \|f - \widetilde{f}\|_n^2
\end{aligned}$$

since $f(v_{i_j}^{(j)}, j \in [m]) = \widetilde{f}(v_{i_j}^{(j)}, j \in [m]) = 0$ if $i_j = 0$ for some $j \in [m]$. Note that

$$\begin{aligned}\frac{(1-r^3)^m}{2\cdot 3^m} - \frac{2^m-1}{3^m} \cdot \sum_{\boldsymbol{\delta}\in\{0,1\}^m\setminus\{\mathbf{1}\}} &\Big[(1-r^3)^{\sum_j \delta_j}(1-r)^{3(m-\sum_j \delta_j)}\Big] \\ = \frac{(1-r)^m}{3^m} \cdot \Bigg[\frac{1}{2}\Big(\frac{1-r^3}{1-r}\Big)^m &- (2^m-1)(1-r)^2 \\ &\cdot \sum_{\boldsymbol{\delta}\in\{0,1\}^m\setminus\{\mathbf{1}\}} \Big(\frac{1-r^3}{1-r}\Big)^{\sum_j \delta_j}(1-r)^{2(m-1-\sum_j \delta_j)}\Bigg].\end{aligned}$$

Therefore, by choosing r close enough to 1, we can show that there exists a positive constant $c_{\rho,m}$ depending on ρ and m such that

$$\big\|\Phi(f) - \Phi(\widetilde{f})\big\|_2 \geq \frac{c_{\rho,m}}{R} \cdot \|f - \widetilde{f}\|_n \tag{C.60}$$

for all $f, \widetilde{f} \in \mathcal{D}_{m,\text{disc}}(R)$. Consequently,

$$\begin{aligned} N(\epsilon, \mathcal{D}_{m,\text{disc}}(R), \|\cdot\|_n) &\leq M(\epsilon, \mathcal{D}_{m,\text{disc}}(R), \|\cdot\|_n) \\ &\leq M\Big(\frac{c_{\rho,m}\epsilon}{R}, \mathcal{D}_m, \|\cdot\|_2\Big) \leq N\Big(\frac{c_{\rho,m}\epsilon}{2R}, \mathcal{D}_m, \|\cdot\|_2\Big).\end{aligned}$$

Here, $M(\epsilon, X, \|\cdot\|_n)$ indicates the ϵ-packing number of a set X under a norm $\|\cdot\|$, the second inequality follows from (C.60), and the first and the last inequalities are from the general theory of packing numbers and covering numbers (see, e.g., [152, Lemma 5.5]). □

C.24 Proof of Lemma 22

Proof of Lemma 22. Note that for every $\mathbf{v} = (v_S, S \subseteq [d], 0 \leq |S| \leq s)$,

$$\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} = \mathbb{E}\Big[\sum_{S,S'}\Big(v_S\prod_{j\in S}x_j^{(1)}\Big)\Big(v_{S'}\prod_{j'\in S'}x_{j'}^{(1)}\Big)\Big] = \mathbb{E}\Big[\Big(\sum_{0\leq|S|\leq s}v_S\prod_{j\in S}x_j^{(1)}\Big)^2\Big] \geq 0.$$

Hence, it suffices to show that $\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} > 0$ for every $\mathbf{v} \neq \mathbf{0}$ in order to prove the lemma. Suppose there exists $\mathbf{v} = (v_S, S \subseteq [d], 0 \leq |S| \leq s)$ such that $\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} = 0$. From the computation above, we can see that

$$\mathbb{E}\Big[\Big(\sum_{0\leq|S|\leq s}v_S\prod_{j\in S}x_j^{(1)}\Big)^2\Big] = 0,$$

which directly implies

$$\mathbb{P}\Big(\sum_{0\leq|S|\leq s}v_S\prod_{j\in S}x_j^{(1)} = 0\Big) = 1. \tag{C.61}$$

For notational convenience, let a_d be the multi-affine function defined by

$$a_d(x_1, \dots, x_d) = \sum_{0 \le |S| \le s} v_S \prod_{j \in S} x_j$$

for $(x_1, \dots, x_d) \in [0,1]^d$. Then, we can restate (C.61) as

$$\mathbb{P}\big(a_d\big(x_1^{(1)}, \dots, x_d^{(1)}\big) = 0\big) = 1.$$

Now, observe that we can write a_d as

$$a_d(x_1, \dots, x_d) = b_{d-1}(x_1, \dots, x_{d-1}) + a_{d-1}(x_1, \dots, x_{d-1}) \cdot x_d \tag{C.62}$$

for some multi-affine functions a_{d-1} and b_{d-1} of $x_1, \dots, x_{d-1}$. We then show that the evaluations of a_{d-1} and b_{d-1} at $\mathbf{x}^{(1)}$ should also be zero almost surely. More precisely, we show that

$$\mathbb{P}\big(a_{d-1}\big(x_1^{(1)}, \dots, x_{d-1}^{(1)}\big) = 0\big) = 1 \quad \text{and} \quad \mathbb{P}\big(b_{d-1}\big(x_1^{(1)}, \dots, x_{d-1}^{(1)}\big) = 0\big) = 1.$$

This is because

$$\begin{aligned} 1 &= \mathbb{P}\big(a_d\big(x_1^{(1)}, \dots, x_d^{(1)}\big) = 0\big) \\ &\le \mathbb{P}\big(a_{d-1}\big(x_1^{(1)}, \dots, x_{d-1}^{(1)}\big) = 0\big) \\ &\quad + \mathbb{P}\big(a_{d-1}\big(x_1^{(1)}, \dots, x_{d-1}^{(1)}\big) \neq 0 \text{ and } x_d^{(1)} = -(b_{d-1}/a_{d-1})\big(x_1^{(1)}, \dots, x_{d-1}^{(1)}\big)\big) \\ &= \mathbb{P}\big(a_{d-1}\big(x_1^{(1)}, \dots, x_{d-1}^{(1)}\big) = 0\big), \end{aligned}$$

where the last equality follows from the assumption that the $\mathbf{x}^{(1)}$ is a continuous random variable with a probability density function.

We can repeat this argument. Suppose we have a multi-affine function of the first i variables $x_1, \dots, x_i$ whose evaluation at $\mathbf{x}^{(1)}$ is zero almost surely. Then, we can represent it as in (C.62) with two multi-affine functions of one less variable, $x_1, \dots, x_{i-1}$ and show as before that the evaluations of those two functions at $\mathbf{x}^{(1)}$ are zero almost surely as well. As the number of variables of multi-affine functions decreases by one at each iteration, we end up with constant functions that are zero almost surely, which means that they are identically zero. If we recursively plug them back into the recurrence relations between multi-affine functions, we can show that a_d is also identically zero, and this proves that the coefficient vector $\mathbf{v}$ of a_d is a zero vector. $\square$

Appendix D

Proofs for Totally Concave Regression

D.1 Proof of Proposition 5

An alternative representation of $\mathcal{F}_{\mathrm{tc}}^{d,s}$, as stated in Lemma 25, is helpful in proving Proposition 5. To describe this representation, we first recall the concept of *entire monotonicity* of functions. Entire monotonicity is a multivariate generalization of univariate monotonicity, defined as follows (see also [46] and references therein). Similar to total concavity in the sense of Popoviciu, entire monotonicity is defined using the divided differences of functions. While total concavity involves divided differences of order $\mathbf{p} = (p_1, \dots, p_d) \in \{0, 1, 2\}^d$ with $\max_j p_j = 2$, entire monotonicity is based on divided differences of order $\mathbf{p} \in \{0, 1\}^d$ with $\max_j p_j = 1$. The proof of Lemma 25 is provided in Appendix D.9.

Definition 9 (Entire Monotonicity)**.** *A real-valued function f on $[0, 1]^d$ is said to be entirely monotone if, for every $(p_1, \dots, p_d)$ with $\max_j p_j = 1$ (i.e., $(p_1, \dots, p_d) \in \{0, 1\}^d \setminus \{\mathbf{0}\}$),*

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & ;\ f \\ v_1^{(d)}, \dots, v_{p_d+1}^{(d)} & \end{bmatrix} \geq 0$$

for every $0 \leq v_1^{(j)} < \cdots < v_{p_j+1}^{(j)} \leq 1$ for $j \in [d]$.

Example 4 ($d = 2$)**.** *A function $f : [0, 1]^2 \to \mathbb{R}$ is entirely monotone if the divided difference of f of order (p_1, p_2) is nonnegative for $(p_1, p_2) = (1, 0), (0, 1), (1, 1)$. The divided difference of f of order $(1, 0)$ is nonnegative if and only if, for every $0 \leq v_1 < v_2 \leq 1$ and $0 \leq w \leq 1$,*

$$\begin{bmatrix} v_1, v_2 & \\ w & ;\ f \end{bmatrix} = \frac{f(v_2, w) - f(v_1, w)}{v_2 - v_1} \geq 0.$$

Thus, the nonnegativity of the divided difference of f of order $(1, 0)$ means that $f(\cdot, x_2)$ is monotone for fixed x_2. Similarly, the nonnegativity of the divided difference of f of order

$(0,1)$ *is equivalent to the monotonicity of* $f(x_1, \cdot)$ *for fixed* x_1. *Also, the divided difference of* f *of order* $(1,1)$ *is nonnegative if and only if*

$$\begin{bmatrix} v_1, v_2 & \\ w_1, w_2 & ; \; f \end{bmatrix} = \frac{f(v_2, w_2) - f(v_1, w_2) - f(v_2, w_1) + f(v_1, w_1)}{(v_2 - v_1)(w_2 - w_1)} \geq 0$$

for every $0 \leq v_1 < v_2 \leq 1$ *and* $0 \leq w_1 < w_2 \leq 1$.

Lemma 25. *The function class* $\mathcal{F}_{tc}^{d,s}$ *consists of all functions of the form*

$$f(x_1, \dots, x_d) = c_0 - \sum_{0 < |S| \leq s} \int_{\prod_{j \in S}[0, x_j]} g_S(t_j, j \in S) \, d(t_j, j \in S) \tag{D.1}$$

for some $c_0 \in \mathbb{R}$ *and some functions* $\{g_S : S \subseteq [d], 0 < |S| \leq s\}$, *where for each* S,

- g_S *is a real-valued function on* $[0,1]^{|S|}$;
- g_S *is entirely monotone;*
- g_S *is right-continuous on* $[0,1]^{|S|}$;
- g_S *is left-continuous at each point* $(x_j, j \in S) \in [0,1]^{|S|} \setminus [0,1)^{|S|}$ *with respect to all the* j^{th} *coordinates where* $x_j = 1$.

In our proof of Proposition 5, the following theorem also plays an important role. This theorem is a version of the fundamental theorem of calculus for interval functions. Recall the definitions of additive interval functions and their derivatives from Appendix C.3. We also say that axis-aligned (closed) rectangles are non-overlapping if no two of them share an interior point. Let R_0 be an axis-aligned rectangle in $\mathbb{R}^m$. For an additive interval function F defined for m-dimensional axis-aligned rectangles contained in R_0, we say that it is absolutely continuous on R_0 if it satisfies the following condition: For every $\epsilon > 0$, there exists $\delta > 0$ such that, for any finite collection of non-overlapping m-dimensional axis-aligned rectangles $R_1, \dots, R_l \subseteq R_0$ with $\sum_{k=1}^{l} |R_k| < \delta$, it holds that $\sum_{k=1}^{l} |F(R_k)| < \epsilon$.

Theorem 27 (Theorem 7.3.3 of [99])**.** *Suppose that* F *is an absolutely continuous interval function on an axis-aligned rectangle* R_0 *in* $\mathbb{R}^m$. *Then,* DF *exists almost everywhere (with respect to the Lebesgue measure) on* R_0, *and*

$$F(R) = \int_R DF(\mathbf{x}) \, d\mathbf{x}$$

for every m*-dimensional axis-aligned rectangle* $R \subseteq R_0$.

As in Appendix C.3, we also use the following notation. For a real-valued function f on $[0,1]^d$ and for each nonempty $S \subseteq [d]$, let $\triangle_S f$ denote the function on intervals, defined by

$$\triangle_S f\left(\prod_{j \in S}[x_j, y_j]\right) = \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S} \delta_j} \cdot f\left(\widetilde{\mathbf{xy}}_{\boldsymbol{\delta}}\right) \tag{D.2}$$

for $0 \le x_j < y_j \le 1$ for $j \in S$. Here, $\widetilde{\mathbf{xy}}_{\boldsymbol{\delta}}$ is the d-dimensional vector for which

$$(\widetilde{\mathbf{xy}}_{\boldsymbol{\delta}})_j = \begin{cases} \delta_j x_j + (1-\delta_j) y_j & \text{if } j \in S \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\triangle_S f$ can be related to the divided differences of f as follows:

$$\triangle_S f\left(\prod_{j \in S}[x_j, y_j]\right) = \left[\prod_{j \in S}(y_j - x_j)\right] \cdot \begin{bmatrix} x_j, y_j, & j \in S & ; & f \\ 0, & j \notin S & & \end{bmatrix}$$

for $0 \le x_j < y_j \le 1$ for $j \in S$. For notational convenience, when $S = [d]$, we omit the subscript and write $\triangle := \triangle_{[d]}$.

Proof of Proposition 5. **Step 1:** $f \in \mathcal{F}^{d,s}_{\mathbf{tc}} \implies f \in \mathcal{F}^{d}_{\mathbf{tcp}}$.

Assume $f \in \mathcal{F}^{d,s}_{\text{tc}}$. We first show that $f \in \mathcal{F}^{d}_{\text{tcp}}$. For this, we need to prove that the divided difference of f of order $\mathbf{p}$ is nonpositive on $[0,1]^d$ for every $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1,2\}^d$ with $\max_j p_j = 2$. The following lemma (proved in Appendix D.10) shows that, in fact, it suffices to verify the nonpositivity of the divided differences of f on some proper subsets of $[0,1]^d$.

Lemma 26. *Let f be a real-valued function on $[0,1]^d$. Then, $f \in \mathcal{F}^d_{tcp}$ if, for each $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1,2\}^d$ with $\max_j p_j = 2$, the divided difference of f of order $\mathbf{p}$ is nonpositive on $N^{(\mathbf{p})} := N_1^{(\mathbf{p})} \times \cdots \times N_d^{(\mathbf{p})}$, where $N_j^{(\mathbf{p})} = [0,1]$ if $p_j \neq 0$ and $N_j^{(\mathbf{p})} = \{0\}$ otherwise, i.e.,*

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & & \\ \vdots & ; & f \\ v_1^{(d)}, \dots, v_{p_d+1}^{(d)} & & \end{bmatrix} \le 0$$

provided that $0 \le v_1^{(j)} < \cdots < v_{p_j+1}^{(j)} \le 1$ if $p_j \neq 0$ and $v_1^{(j)} = 0$ otherwise.

Without loss of generality, we assume that $\mathbf{p} = (2, \dots, 2, 1, \dots, 1, 0, \dots, 0)$ where $|\{j : p_j \neq 0\}| = m_2$ and $|\{j : p_j = 2\}| = m_1$. Also, for notational convenience, we let $S_{\mathbf{p}} = [m_2]$. By Lemma 25, there exist $c_0 \in \mathbb{R}$ and functions g_S, $0 < |S| \le s$ satisfying the conditions in Lemma 25 such that f can be expressed as in (D.1). For every $0 \le v_1^{(j)} < \cdots < v_{p_j+1}^{(j)} \le 1$, $j \in [m_2]$, we can thus represent the divided difference of f on $N^{(\mathbf{p})}$ as follows:

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & & \\ \vdots & & \\ v_1^{(m_2)}, \dots, v_{p_{m_2}+1}^{(m_2)} & ; & f \\ 0 & & \\ \vdots & & \\ 0 & & \end{bmatrix}$$

$$
\begin{aligned}
&= \sum_{i_1=1}^{3} \cdots \sum_{i_{m_1}=1}^{3} \sum_{i_{m_1+1}=1}^{2} \cdots \sum_{i_{m_2}=1}^{2} \frac{1}{\prod_{l_1 \neq i_1} (v_{i_1}^{(1)} - v_{l_1}^{(1)})} \times \cdots \times \frac{1}{\prod_{l_{m_1} \neq i_{m_1}} (v_{i_{m_1}}^{(m_1)} - v_{l_{m_1}}^{(m_1)})} \\
&\qquad \cdot \frac{1}{\prod_{j=m_1+1}^{m_2} (v_2^{(j)} - v_1^{(j)})} \cdot (-1)^{\sum_{j=m_1+1}^{m_2} i_j} \cdot f\big(v_{i_1}^{(1)}, \ldots, v_{i_{m_2}}^{(m_2)}, 0, \ldots, 0\big) \\
&= \frac{1}{\prod_{j=1}^{m_1} (v_3^{(j)} - v_1^{(j)})} \cdot \sum_{\boldsymbol{\delta} \in \{0,1\}^{m_1}} (-1)^{\sum_{j=1}^{m_1} \delta_j} \cdot \frac{1}{\prod_{j=1}^{m_1} (v_{3-\delta_j}^{(j)} - v_{2-\delta_j}^{(j)})} \cdot \frac{1}{\prod_{j=m_1+1}^{m_2} (v_2^{(j)} - v_1^{(j)})} \\
&\qquad \cdot \sum_{i_1 = 2-\delta_1}^{3-\delta_1} \cdots \sum_{i_{m_1} = 2-\delta_{m_1}}^{3-\delta_{m_1}} \sum_{i_{m_1+1}=1}^{2} \cdots \sum_{i_{m_2}=1}^{2} \\
&\qquad\qquad (-1)^{\sum_{j=1}^{m_1} (i_j + \delta_j - 1)} \cdot (-1)^{\sum_{j=m_1+1}^{m_2} i_j} f\big(v_{i_1}^{(1)}, \ldots, v_{i_{m_2}}^{(m_2)}, 0, \ldots, 0\big) \\
&= \frac{1}{\prod_{j=1}^{m_1} (v_3^{(j)} - v_1^{(j)})} \cdot \frac{1}{\prod_{j=m_1+1}^{m_2} (v_2^{(j)} - v_1^{(j)})} \cdot \sum_{\boldsymbol{\delta} \in \{0,1\}^{m_1}} (-1)^{\sum_{j=1}^{m_1} \delta_j} \cdot \frac{1}{\prod_{j=1}^{m_1} (v_{3-\delta_j}^{(j)} - v_{2-\delta_j}^{(j)})} \\
&\qquad \cdot \triangle_{S_{\mathbf{p}}} f \Bigg(\prod_{j=1}^{m_1} \big[v_{2-\delta_j}^{(j)}, v_{3-\delta_j}^{(j)}\big] \times \prod_{j=m_1+1}^{m_2} \big[v_1^{(j)}, v_2^{(j)}\big] \Bigg) \\
&= -\frac{1}{\prod_{j=1}^{m_1} (v_3^{(j)} - v_1^{(j)})} \cdot \frac{1}{\prod_{j=m_1+1}^{m_2} (v_2^{(j)} - v_1^{(j)})} \cdot \sum_{\boldsymbol{\delta} \in \{0,1\}^{m_1}} (-1)^{\sum_{j=1}^{m_1} \delta_j} \cdot \frac{1}{\prod_{j=1}^{m_1} (v_{3-\delta_j}^{(j)} - v_{2-\delta_j}^{(j)})} \\
&\qquad \cdot \int_{v_{2-\delta_1}^{(1)}}^{v_{3-\delta_1}^{(1)}} \cdots \int_{v_{2-\delta_{m_1}}^{(m_1)}}^{v_{3-\delta_{m_1}}^{(m_1)}} \int_{v_1^{(m_1+1)}}^{v_2^{(m_1+1)}} \cdots \int_{v_1^{(m_2)}}^{v_2^{(m_2)}} g_{S_{\mathbf{p}}}(t_1, \ldots, t_{m_2}) \, dt_{m_2} \cdots dt_1.
\end{aligned}
$$

Also, since $g_{S_{\mathbf{p}}}$ is entirely monotone, we have

$$
\begin{aligned}
0 &\leq \int_{v_2^{(1)}}^{v_3^{(1)}} \int_{v_1^{(1)}}^{v_2^{(1)}} \cdots \int_{v_2^{(m_1)}}^{v_3^{(m_1)}} \int_{v_1^{(m_1)}}^{v_2^{(m_1)}} \int_{v_1^{(m_1+1)}}^{v_2^{(m_1+1)}} \cdots \int_{v_1^{(m_2)}}^{v_2^{(m_2)}} \\
&\qquad \prod_{j=1}^{m_1} (w_j - u_j) \cdot \begin{bmatrix} u_1, w_1 & \\ \vdots & \\ u_{m_1}, w_{m_1} & ; \ g_{S_{\mathbf{p}}} \\ t_{m_1+1} & \\ \vdots & \\ t_{m_2} & \end{bmatrix} dt_{m_2} \cdots dt_{m_1+1} du_{m_1} dw_{m_1} \cdots du_1 dw_1 \\
&= \prod_{j=1}^{m_1} \big[\big(v_3^{(j)} - v_2^{(j)}\big)\big(v_2^{(j)} - v_1^{(j)}\big) \big] \cdot \sum_{\boldsymbol{\delta} \in \{0,1\}^{m_1}} (-1)^{\sum_{j=1}^{m_1} \delta_j} \cdot \frac{1}{\prod_{j=1}^{m_1} (v_{3-\delta_j}^{(j)} - v_{2-\delta_j}^{(j)})} \\
&\qquad \cdot \int_{v_{2-\delta_1}^{(1)}}^{v_{3-\delta_1}^{(1)}} \cdots \int_{v_{2-\delta_{m_1}}^{(m_1)}}^{v_{3-\delta_{m_1}}^{(m_1)}} \int_{v_1^{(m_1+1)}}^{v_2^{(m_1+1)}} \cdots \int_{v_1^{(m_2)}}^{v_2^{(m_2)}} g_{S_{\mathbf{p}}}(t_1, \ldots, t_{m_2}) \, dt_{m_2} \cdots dt_1.
\end{aligned}
$$

Combining these results, we can show that the divided difference of order $\mathbf{p}$ is nonpositive on $N^{(\mathbf{p})} = [0,1]^{m_2} \times \{0\}^{d-m_2}$ when $\mathbf{p} = (2,\dots,2,1,\dots,1,0,\dots,0)$. Through the same argument, we can prove that this is also true for every $\mathbf{p} \in \{0,1,2\}^d$ with $\max_j p_j = 2$. Due to Lemma 26, we can therefore conclude that $f \in \mathcal{F}^d_{\mathrm{tcp}}$.

Step 2: $f \in \mathcal{F}^{d,s}_{\mathbf{tc}} \implies$ (4.14) and (4.15) hold for all $S \neq \emptyset$ and (4.16) holds for all S with $|S| > s$.

We now show that f satisfies (4.14) and (4.15) for all nonempty subsets $S \subseteq [d]$ and (4.16) for all subsets S with $|S| > s$. It can be readily checked that (4.16) holds for all subsets S with $|S| > s$. Also, for each subset $S \subseteq [d]$ with $0 < |S| \le s$, since g_S is entirely monotone, we have that

$$\begin{bmatrix} 0, t_j, & j \in S \\ 0, & j \notin S \end{bmatrix} ; \; f \Bigg] = -\frac{1}{\prod_{j \in S} t_j} \int_{\prod_{j\in S}(0,t_j]} g_S(s_j, j \in S)\, d(s_j, j \in S) \le -g_S(0, j \in S)$$

for every $t_j > 0$ for $j \in S$, and that

$$\begin{bmatrix} t_j, 1, & j \in S \\ 0, & j \notin S \end{bmatrix} ; \; f \Bigg] = -\frac{1}{\prod_{j \in S}(1 - t_j)} \int_{\prod_{j\in S}(t_j,1]} g_S(s_j, j \in S)\, d(s_j, j \in S) \ge -g_S(1, j \in S)$$

for every $t_j < 1$ for $j \in S$. The conditions (4.14) and (4.15) follow directly from these results.

Step 3: $f \in \mathcal{F}^d_{\mathbf{tcp}}$ and (4.14), (4.15) hold for all $S \neq \emptyset \implies f \in \mathcal{F}^d_{\mathbf{tc}}$.

Next, we prove that, for every $f \in \mathcal{F}^d_{\mathrm{tcp}}$ satisfying (4.14) and (4.15) for all nonempty subsets $S \subseteq [d]$, it holds that $f \in \mathcal{F}^d_{\mathrm{tc}}$. Observe that

$$f(x_1,\dots,x_d) = f(0,\dots,0) + \sum_{\emptyset \neq S \subseteq [d]} \triangle_S f\left(\prod_{j\in S}[0,x_j]\right)$$

for every $(x_1,\dots,x_d) \in [0,1]^d$. Hence, by Lemma 25, it suffices to show that, for each nonempty S, there exists a function g_S satisfying the conditions of Lemma 25 such that

$$\triangle_S f\left(\prod_{j\in S}[0,x_j]\right) = -\int_{\prod_{j\in S}[0,x_j]} g_S(t_j, j\in S)\, d(t_j, j \in S) \tag{D.3}$$

for every $(x_j, j \in S) \in [0,1]^{|S|}$. Here, we prove this claim only for the case $S = [d]$, but our argument can be readily adapted to other nonempty subsets S of $[d]$.

Step 3-1: $\triangle f$ is absolutely continuous.

Recall that $\triangle = \triangle_{[d]}$. We first show that $\triangle f$ is an absolutely continuous interval function on $[0,1]^d$. To show this, we use the following lemma, which we prove in Appendix D.11.

Lemma 27. *Suppose f is a real-valued function on $[0,1]^d$ whose divided difference of order $\mathbf{p} = (p_1,\dots,p_d)$ is nonpositive for $\mathbf{p} = (2,1,\dots,1),(1,2,1,\dots,1),\dots,(1,\dots,1,2)$. Then,*

$$\begin{bmatrix} v_1^{(1)}, v_2^{(1)} & \\ \vdots & ; \; f \\ v_1^{(d)}, v_2^{(d)} & \end{bmatrix} \ge \begin{bmatrix} v_1^{(1)}, v_3^{(1)} & \\ \vdots & ; \; f \\ v_1^{(d)}, v_3^{(d)} & \end{bmatrix}$$

if $0 \leq v_1^{(j)} < v_2^{(j)} \leq v_3^{(j)} \leq 1$ *for* $j \in [d]$, *and*

$$\begin{bmatrix} v_1^{(1)}, v_3^{(1)} & & \\ \vdots & ; & f \\ v_1^{(d)}, v_3^{(d)} & & \end{bmatrix} \geq \begin{bmatrix} v_2^{(1)}, v_3^{(1)} & & \\ \vdots & ; & f \\ v_2^{(d)}, v_3^{(d)} & & \end{bmatrix}$$

if $0 \leq v_1^{(j)} \leq v_2^{(j)} < v_3^{(j)} \leq 1$ *for* $j \in [d]$. *Also, it follows that*

$$\begin{bmatrix} v_1^{(1)}, v_2^{(1)} & & \\ \vdots & ; & f \\ v_1^{(d)}, v_2^{(d)} & & \end{bmatrix} \geq \begin{bmatrix} w_1^{(1)}, w_2^{(1)} & & \\ \vdots & ; & f \\ w_1^{(d)}, w_2^{(d)} & & \end{bmatrix}$$

if $0 \leq v_1^{(j)} \leq w_1^{(j)} \leq 1$ *and* $0 \leq v_2^{(j)} \leq w_2^{(j)} \leq 1$ *for* $j \in [d]$.

Along with the conditions (4.14) and (4.15), Lemma 27 implies that, for every $0 \leq v_j^{(1)} < v_j^{(2)} \leq 1$, $j \in [d]$,

$$\left| \frac{1}{\prod_{j=1}^d (v_j^{(2)} - v_j^{(1)})} \cdot \triangle f\left(\prod_{j=1}^d \left[v_j^{(1)}, v_j^{(2)}\right] \right) \right| = \left| \begin{bmatrix} v_1^{(1)}, v_2^{(1)} & & \\ \vdots & ; & f \\ v_1^{(d)}, v_2^{(d)} & & \end{bmatrix} \right| \tag{D.4}$$
$$\leq \max\left(\left| \lim_{t \to 1-} \begin{bmatrix} t, 1 & & \\ \vdots & ; & f \\ t, 1 & & \end{bmatrix} \right|, \left| \lim_{t \to 0+} \begin{bmatrix} 0, t & & \\ \vdots & ; & f \\ 0, t & & \end{bmatrix} \right| \right) < \infty,$$

from which the absolute continuity of $\triangle f$ follows directly. Therefore, by Theorem 27, $D\triangle f$ exists almost everywhere (w.r.t. the Lebesgue measure) on $[0, 1]^d$, and we have

$$\triangle f\left(\prod_{j=1}^d [0, x_j] \right) = \int_{\prod_{j=1}^d [0, x_j]} D\triangle f(t_1, \dots, t_d)\, d(t_1, \dots, t_d) \tag{D.5}$$

for every $(x_1, \dots, x_d) \in [0, 1]^d$. Furthermore, using Lemma 27 again, we can show that

$$D\triangle f(x_1, \dots, x_d) \geq D\triangle f(x_1', \dots, x_d') \tag{D.6}$$

whenever $x_j \leq x_j'$ for $j \in [d]$, and they exist.

Step 3-2: Definition of g.

Now, we define a function g on $[0, 1]^d$ as follows. For $\mathbf{x} = (x_1, \dots, x_d) \in [0, 1)^d$, let

$$g(x_1, \dots, x_d) = \inf\left\{ -D\triangle f(s_1, \dots, s_d) : s_j > x_j \text{ and } D\triangle f(s_1, \dots, s_d) \text{ exists} \right\}. \tag{D.7}$$

For $\mathbf{x} = (x_1, \dots, x_d) \in [0, 1]^d \setminus [0, 1)^d$, the value of g at $\mathbf{x}$ is determined sequentially as follows:

$$g(1, x_2, \dots, x_d) = \lim_{t_1 \uparrow 1} g(t_1, x_2, \dots, x_d) \quad \text{for } x_2, \dots, x_d < 1,$$

$$
\begin{aligned}
&g(x_1, 1, x_3, \dots, x_d) = \lim_{t_2 \uparrow 1} g(x_1, t_2, x_3, \dots, x_d) \ \text{ for } x_1, x_3, \dots, x_d < 1, \ \dots \ , \\
&g(x_1, \dots, x_{d-1}, 1) = \lim_{t_d \uparrow 1} g(x_1, \dots, x_{d-1}, t_d) \ \text{ for } x_1, \dots, x_{d-1} < 1, \\
&g(1, 1, x_3, \dots, x_d) = \lim_{t_1 \uparrow 1} g(t_1, 1, x_3, \dots, x_d) = \lim_{t_2 \uparrow 1} g(1, t_2, x_3, \dots, x_d) \ \text{ for } x_3, \dots, x_d < 1, \ \dots \ , \\
&g(x_1, \dots, x_{d-2}, 1, 1) = \lim_{t_{d-1} \uparrow 1} g(x_1, \dots, x_{d-2}, t_{d-1}, 1) = \lim_{t_d \uparrow 1} g(x_1, \dots, x_{d-2}, 1, t_d) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } x_1, \dots, x_{d-2} < 1, \ \dots \ , \\
&g(1, \dots, 1) = \lim_{t_1 \uparrow 1} g(t_1, 1, \dots, 1) = \dots = \lim_{t_d \uparrow 1} g(1, \dots, 1, t_d).
\end{aligned}
$$

If we prove that (a) g is well-defined, (b) g satisfies the conditions of Lemma 25, and (c) $g = -D\triangle f$ almost everywhere (w.r.t. the Lebesgue measure), then by (D.5), we can guarantee that g is the desired function for the claim (D.3) when $S = [d]$.

Step 3-3: Well-definedness of g.

We prove that g is well-defined at every point $\mathbf{x} = (x_1, \dots, x_d) \in [0,1]^d$ by induction on $|\{j : x_j = 1\}|$. From (D.4) and (D.7), it follows that

$$\left| \sup_{(x_1, \dots, x_d) \in [0,1)^d} g(x_1, \dots, x_d) \right| < \infty.$$

Moreover, by (D.7), $g(x_1, \dots, x_d) \leq g(x'_1, \dots, x'_d)$ for every $x_j \leq x'_j < 1$, $j \in [d]$. These boundedness and monotonicity of g ensure that, for fixed $x_2, \dots, x_d < 1$, the limit

$$\lim_{t_1 \uparrow 1} g(t_1, x_2, \dots, x_d)$$

exists, which means that $g(1, x_2, \dots, x_d)$ is well-defined. Through the same argument, we can show that g is well-defined at every $\mathbf{x} \in [0,1]^d$ with $|\{j : x_j = 1\}| = 1$.

We now assume that g is well-defined at every $\mathbf{x} \in [0,1]^d$ with $|\{j : x_j = 1\}| \leq m - 1$ and show that it is also well-defined at every $\mathbf{x}$ with $|\{j : x_j = 1\}| = m$. Without loss of generality, assume $\mathbf{x} = (1, \dots, 1, x_{m+1}, \dots, x_d)$, where $x_{m+1}, \dots, x_d < 1$. To show that g is well-defined at $\mathbf{x}$, we need to verify that

$$\lim_{t_1 \uparrow 1} g(t_1, 1, \dots, 1, x_{m+1}, \dots, x_d), \dots, \lim_{t_m \uparrow 1} g(1, \dots, 1, t_m, x_{m+1}, \dots, x_d)$$

exist and coincide. To establish this, it suffices to prove that

$$\lim_{t_m \uparrow 1} g(1, \dots, 1, t_m, x_{m+1}, \dots, x_d)$$

exists and satisfies

$$\lim_{t_m \uparrow 1} g(1, \dots, 1, t_m, x_{m+1}, \dots, x_d) = \lim_{n \to \infty} g\Big(1 - \frac{1}{n}, \dots, 1 - \frac{1}{n}, x_{m+1}, \dots, x_d\Big).$$

The existence of the limit follows from the boundedness of g and the monotonicity:

$$\begin{aligned} g(1,\dots,1,t_m,x_{m+1},\dots,x_d) &= \lim_{t_1\uparrow 1}\cdots\lim_{t_{m-1}\uparrow 1} g(t_1,\dots,t_{m-1},t_m,x_{m+1},\dots,x_d) \\ &\le \lim_{t_1\uparrow 1}\cdots\lim_{t_{m-1}\uparrow 1} g(t_1,\dots,t_{m-1},t'_m,x_{m+1},\dots,x_d) = g(1,\dots,1,t'_m,x_{m+1},\dots,x_d) \end{aligned}$$

for every $t_m < t'_m < 1$. Next, note that

$$\begin{aligned} \lim_{n\to\infty} g\Big(1-\frac{1}{n},\dots,1-\frac{1}{n},x_{m+1},\dots,x_d\Big) &\le \lim_{n\to\infty} g\Big(1,1-\frac{1}{n},\dots,1-\frac{1}{n},x_{m+1},\dots,x_d\Big) \\ \le \cdots \le \lim_{n\to\infty} g\Big(1,\dots,1,1-\frac{1}{n},x_{m+1},\dots,x_d\Big) &= \lim_{t_m\uparrow 1} g(1,\dots,1,t_m,x_{m+1},\dots,x_d). \end{aligned}$$

Thus, it remains to prove that

$$\lim_{n\to\infty} g\Big(1-\frac{1}{n},\dots,1-\frac{1}{n},x_{m+1},\dots,x_d\Big) \ge \lim_{t_m\uparrow 1} g(1,\dots,1,t_m,x_{m+1},\dots,x_d). \tag{D.8}$$

Fix $\epsilon > 0$. Then, there exists $s_m < 1$ such that

$$g(1,\dots,1,s_m,x_{m+1},\dots,x_d) > \lim_{t_m\uparrow 1} g(1,\dots,1,t_m,x_{m+1},\dots,x_d) - \frac{\epsilon}{m}.$$

Similarly, there exists $s_{m-1} < 1$ such that

$$\begin{aligned} g(1,\dots,1,s_{m-1},s_m,x_{m+1},\dots,x_d) &> g(1,\dots,1,1,s_m,x_{m+1},\dots,x_d) - \frac{\epsilon}{m} \\ &> \lim_{t_m\uparrow 1} g(1,\dots,1,1,t_m,x_{m+1},\dots,x_d) - \frac{2\epsilon}{m}. \end{aligned}$$

Repeating this argument, we can find $s_1,\dots,s_m < 1$ for which

$$g(s_1,\dots,s_m,x_{m+1},\dots,x_d) > \lim_{t_m\uparrow 1} g(1,\dots,1,t_m,x_{m+1},\dots,x_d) - \epsilon.$$

This implies that

$$\begin{aligned} \lim_{n\to\infty} g\Big(1-\frac{1}{n},\dots,1-\frac{1}{n},x_{m+1},\dots,x_d\Big) &\ge g(s_1,\dots,s_m,x_{m+1},\dots,x_d) \\ &> \lim_{t_m\uparrow 1} g(1,\dots,1,t_m,x_{m+1},\dots,x_d) - \epsilon. \end{aligned}$$

Since $\epsilon > 0$ can be arbitrarily small, this proves (D.8), and therefore g is well-defined at $\mathbf{x}$. By induction, we can see that g is well-defined at every point in $[0,1]^d$.

Step 3-4: Entire monotonicity of g.

Next, we show that g is entirely monotone. To prove this, we use the following lemma, which derives the nonpositivity of the divided difference of $D\triangle f$ of order $\mathbf{p} = (p_1,\dots,p_d)$ from the nonpositivity of the divided difference of f of order $(p_1+1,\dots,p_d+1)$ for $\mathbf{p} \in \{0,1\}^d \setminus \{\mathbf{0}\}$. The proof of Lemma 28 is deferred to Appendix D.12.

Lemma 28. *Fix* $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1\}^d \setminus \{\mathbf{0}\}$. *Suppose* f *is a real-valued function on* $[0,1]^d$ *whose divided difference of order* $(p_1+1, \dots, p_d+1)$ *is nonpositive. Then, for every* $0 \le v_1^{(j)} < \dots < v_{p_j+1}^{(j)} \le 1$, $j \in [d]$, *we have*

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & ;\ D\triangle f \\ v_1^{(d)}, \dots, v_{p_d+1}^{(d)} & \end{bmatrix} \le 0,$$

provided that $D\triangle f(v_{i_1}^{(1)}, \dots, v_{i_d}^{(d)})$ *exists for every* $i_j = 1, \dots, p_j+1$, $j \in [d]$.

Fix $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1\}^d \setminus \{\mathbf{0}\}$. We prove that the divided difference of g of order $\mathbf{p}$ with respect to the points $(v_{i_j}^{(j)}, 1 \le i_j \le p_j + 1, 1 \le j \le d)$ is nonnegative for every $0 \le v_1^{(j)} < \dots < v_{p_j+1}^{(j)} \le 1$, $j \in [d]$. We first consider the case where $v_{p_j+1}^{(j)} < 1$ for all $j \in [d]$. Since $D\triangle f$ exists almost everywhere, there exists a sequence of positive real vectors $\{(\delta_n^{(1)}, \dots, \delta_n^{(d)}) : n \in \mathbb{N}\}$ such that $\lim_{n\to\infty} \delta_n^{(j)} = 0$ for each $j \in [d]$, and $D\triangle f(v_{i_1}^{(1)} + \delta_n^{(1)}, \dots, v_{i_d}^{(d)} + \delta_n^{(d)})$ exists for every $i_j = 1, \dots, p_j+1$, $j \in [d]$ and for every $n \in \mathbb{N}$. From the monotonicity (D.6) of $D\triangle f$ and the definition (D.7) of g, it follows that

$$g\big(v_{i_1}^{(1)}, \dots, v_{i_d}^{(d)}\big) = -\lim_{n\to\infty} D\triangle f\big(v_{i_1}^{(1)} + \delta_n^{(1)}, \dots, v_{i_d}^{(d)} + \delta_n^{(d)}\big)$$

for every $i_j = 1, \dots, p_j + 1$, $j \in [d]$. This implies that

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & ;\ g \\ v_1^{(d)}, \dots, v_{p_d+1}^{(d)} & \end{bmatrix} = -\lim_{n\to\infty} \begin{bmatrix} v_1^{(1)} + \delta_n^{(1)}, \dots, v_{p_1+1}^{(1)} + \delta_n^{(1)} & \\ \vdots & ;\ D\triangle f \\ v_1^{(d)} + \delta_n^{(d)}, \dots, v_{p_d+1}^{(d)} + \delta_n^{(d)} & \end{bmatrix}.$$

Since $\max_j(p_j+1) = 2$, the divided difference of f of order $(p_1+1, \dots, p_d+1)$ is nonpositive. By Lemma 28, we thus have

$$\begin{bmatrix} v_1^{(1)} + \delta_n^{(1)}, \dots, v_{p_1+1}^{(1)} + \delta_n^{(1)} & \\ \vdots & ;\ D\triangle f \\ v_1^{(d)} + \delta_n^{(d)}, \dots, v_{p_d+1}^{(d)} + \delta_n^{(d)} & \end{bmatrix} \le 0$$

for every $n \in \mathbb{N}$. This proves that the divided difference of g of order $\mathbf{p}$ with respect to the points $(v_{i_j}^{(j)}, 1 \le i_j \le p_j + 1, 1 \le j \le d)$ is nonnegative.

Now, we consider the case where $v_{p_j+1}^{(j)} = 1$ for some $j \in [d]$. For notational convenience, we assume that $v_{p_j+1}^{(j)} = 1$ for $j \le m$ and $v_{p_j+1}^{(j)} < 1$ for $j > m$. Then, by the definition (D.7) of g and the result established in the previous paragraph, we have

$$\begin{bmatrix} v_1^{(1)}, \ldots, v_{p_1+1}^{(1)} & \\ \vdots & ;\ g \\ v_1^{(d)}, \ldots, v_{p_d+1}^{(d)} & \end{bmatrix} = \lim_{t_1 \uparrow 1} \cdots \lim_{t_m \uparrow 1} \begin{bmatrix} v_1^{(1)}, \ldots, v_{p_1}^{(1)}, t_1 & \\ \vdots & \\ v_1^{(m)}, \ldots, v_{p_m}^{(m)}, t_m & ;\ g \\ v_1^{(m+1)}, \ldots, v_{p_{m+1}+1}^{(m+1)} & \\ \vdots & \\ v_1^{(d)}, \ldots, v_{p_d+1}^{(d)} & \end{bmatrix} \geq 0,$$

which means that the divided difference of g of order $\mathbf{p}$ with respect to the points $(v_{i_j}^{(j)}, 1 \leq i_j \leq p_j + 1, 1 \leq j \leq d)$ is nonnegative in this case as well. Thus, we can conclude that g is entirely monotone.

Step 3-5: Coordinate-wise right and left continuity of g.

It is clear from the definition that g is left-continuous at each point $(x_j, j \in S) \in [0,1]^{|S|} \setminus [0,1)^{|S|}$ with respect to all the j^{th} coordinates where $x_j = 1$. Hence, to show that g satisfies the conditions of Lemma 25, it remains to verify the right-continuity of g. The right-continuity of g on $[0,1)^d$ is also clear from the definition of g. We prove that g is also right-continuous at every point $\mathbf{x} = (x_1, \ldots, x_d) \in [0,1]^d \setminus [0,1)^d$ by induction on $|\{j : x_j = 1\}|$.

We assume that g is right-continuous at every $\mathbf{x} \in [0,1]^d \setminus [0,1)^d$ with $|\{j : x_j = 1\}| \leq m-1$ and show that it is also right-continuous at every $\mathbf{x} \in [0,1]^d \setminus [0,1)^d$ with $|\{j : x_j = 1\}| = m$. Without loss of generality, let $\mathbf{x} = (1, \ldots, 1, x_{m+1}, \ldots, x_d)$, where $x_{m+1}, \ldots, x_d < 1$. Suppose, for contradiction, g is not right-continuous at $\mathbf{x}$. Then, without loss of generality, we can assume that

$$g(1, \ldots, 1, x_{m+1}, \ldots, x_d) < \lim_{t_{m+1} \downarrow x_{m+1}} g(1, \ldots, 1, t_{m+1}, x_{m+2}, \ldots, x_d).$$

Let $\epsilon = \lim_{t_{m+1} \downarrow x_{m+1}} g(1, \ldots, 1, t_{m+1}, x_{m+2}, \ldots, x_d) - g(1, \ldots, 1, x_{m+1}, \ldots, x_d) > 0$ and choose any $s_{m+1} > x_{m+1}$. Since

$$g(1, \ldots, 1, 1, s_{m+1}, x_{m+2}, \ldots, x_d) = \lim_{t_m \uparrow 1} g(1, \ldots, 1, t_m, s_{m+1}, x_{m+2}, \ldots, x_d),$$

there exists $s_m < 1$ such that

$$g(1, \ldots, 1, s_m, s_{m+1}, x_{m+2}, \ldots, x_d) > g(1, \ldots, 1, 1, s_{m+1}, x_{m+2}, \ldots, x_d) - \frac{\epsilon}{2}.$$

For every $t_{m+1} \in (x_{m+1}, s_{m+1})$, the entire monotonicity of g implies that

$$\begin{aligned} g(1, \ldots, 1, s_m, t_{m+1}, x_{m+2}, \ldots, x_d) &\geq g(1, \ldots, 1, 1, t_{m+1}, x_{m+2}, \ldots, x_d) \\ &\quad + \big(g(1, \ldots, 1, s_m, s_{m+1}, x_{m+2}, \ldots, x_d) - g(1, \ldots, 1, 1, s_{m+1}, x_{m+2}, \ldots, x_d)\big) \\ &\geq g(1, \ldots, 1, 1, t_{m+1}, x_{m+2}, \ldots, x_d) - \frac{\epsilon}{2}. \end{aligned}$$

By the right-continuity of g at $(1, \ldots, 1, s_m, x_{m+1}, \ldots, x_d)$ with respect to the $(m+1)^{\text{th}}$ coordinate, we have

$$\begin{aligned}
g(1, \ldots, 1, s_m, x_{m+1}, \ldots, x_d) &= \lim_{t_{m+1} \downarrow x_{m+1}} g(1, \ldots, 1, s_m, t_{m+1}, x_{m+2}, \ldots, x_d) \\
&\geq \lim_{t_{m+1} \downarrow x_{m+1}} g(1, \ldots, 1, 1, t_{m+1}, x_{m+2}, \ldots, x_d) - \frac{\epsilon}{2} = g(1, \ldots, 1, 1, x_{m+1}, x_{m+2}, \ldots, x_d) + \frac{\epsilon}{2} \\
&\geq g(1, \ldots, 1, s_m, x_{m+1}, \ldots, x_d) + \frac{\epsilon}{2},
\end{aligned}$$

which leads to a contradiction. Hence, g is right-continuous at $\mathbf{x}$. By induction, we can conclude that g is right-continuous on the entire domain $[0, 1]^d$.

Step 3-6: $g = -D\triangle f$ almost everywhere.

Lastly, we show that $g = -D\triangle f$ almost everywhere. Consider the set

$$A = \big\{(x_1, \ldots, x_d) \in [0, 1)^d : D\triangle f(x_1, \ldots, x_d) \text{ exists, } g(x_1, \ldots, x_d) \neq -D\triangle f(x_1, \ldots, x_d)\big\}.$$

It suffices to show that $m(A) = 0$, where m denotes the Lebesgue measure. First, observe that if $x_j < x'_j$, $j \in [d]$, and $D\triangle f(x_1, \ldots, x_d)$, $D\triangle f(x'_1, \ldots, x'_d)$ exist, then by (D.6) and (D.7),

$$-D\triangle f(x_1, \ldots, x_d) \leq g(x_1, \ldots, x_d) \leq -D\triangle f(x'_1, \ldots, x'_d) \leq g(x'_1, \ldots, x'_d).$$

This implies that, for each $\mathbf{c} = (c_1, \ldots, c_d) \in \mathbb{R}^d$, there exists an injective map from the set

$$A_{\mathbf{c}} := \big\{u \in \mathbb{R} : u \cdot (1, \ldots, 1) + \mathbf{c} \in A\big\}$$

to $\mathbb{Q}$, the set of rational numbers, from which it follows that $A_{\mathbf{c}}$ is at most countable. Let $\mathbf{1} = (1, \ldots, 1)$ and let $\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)$, $j \in [d]$ be the standard basis vectors of $\mathbb{R}^d$. Also, let $\mathbf{P}$ be a $d \times d$ orthogonal matrix satisfying

$$\mathbf{P}\Big(\frac{1}{\sqrt{d}} \cdot \mathbf{1}\Big) = \mathbf{e}_1$$

and let $\phi_{\mathbf{P}}$ be the corresponding linear map defined by $\phi_{\mathbf{P}}(\mathbf{x}) = \mathbf{P}\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^d$. Then, we can derive that

$$\begin{aligned}
m(A) = m(\phi_{\mathbf{P}}(A)) &= \int \cdots \int \mathbf{1}\big((y_1, \ldots, y_d) \in \phi_{\mathbf{P}}(A)\big)\, dy_1 \cdots dy_d \\
&= \int \cdots \int \mathbf{1}\Big(\sum_{j=1}^d y_j \mathbf{e}_j \in \phi_{\mathbf{P}}(A)\Big)\, dy_1 \cdots dy_d \\
&= \int \cdots \int \mathbf{1}\Big(\sum_{j=1}^d y_j \phi_{\mathbf{P}}^{-1}(\mathbf{e}_j) \in A\Big)\, dy_1 \cdots dy_d
\end{aligned}$$

$$= \int \cdots \int \mathbf{1}\bigg(\frac{y_1}{\sqrt{d}} \cdot \mathbf{1} + \sum_{j=2}^{d} y_j \phi_{\mathbf{P}}^{-1}(\mathbf{e}_j) \in A\bigg)\, dy_1 \cdots dy_d = \int \cdots \int 0\, dy_2 \cdots dy_d = 0,$$

where the second-to-last inequality is from the fact that

$$\mathbf{1}\bigg(y_1 \in \mathbb{R} : \frac{y_1}{\sqrt{d}} \cdot \mathbf{1} + \sum_{j=2}^{d} y_j \phi_{\mathbf{P}}^{-1}(\mathbf{e}_j) \in A\bigg)$$

is at most countable for fixed $y_2, \ldots, y_d \in \mathbb{R}$.

To sum up, we showed that g satisfies the conditions of Lemma 25 and $g = -D\triangle f$ almost everywhere. By (D.5), it follows that g is the desired function for the claim (D.3) when $S = [d]$. Using similar arguments, we can find a function g_S for the claim (D.3) for each nonempty subset S. Consequently, Lemma 25 proves that $f \in \mathcal{F}_{\text{tc}}^{d}$.

Step 4: $f \in \mathcal{F}_{\mathbf{tc}}^{d}$ and (4.16) holds for all S with $|S| > s \implies f \in \mathcal{F}_{\mathbf{tc}}^{d,s}$.

We wrap up the proof by showing that if we further assume that f satisfies (4.16) for every subset S with $|S| > s$, then $f \in \mathcal{F}_{\text{tc}}^{d,s}$. In this case, it follows directly that $\triangle_S f = 0$ for every subset S with $|S| > s$. Thus, we can express f as

$$f(x_1, \ldots, x_d) = f(0, \ldots, 0) + \sum_{\emptyset \neq S \subseteq [d]} \triangle_S f\bigg(\prod_{j \in S}[0, x_j]\bigg) = f(0, \ldots, 0) + \sum_{0 < |S| \leq s} \triangle_S f\bigg(\prod_{j \in S}[0, x_j]\bigg)$$

for $(x_1, \ldots, x_d) \in [0,1]^d$. By using g_S satisfying the claim (D.3) and Lemma 25 again, we can therefore conclude that $f \in \mathcal{F}_{\text{tc}}^{d,s}$. $\square$

D.2 Proof of Lemma 3

Proof of Lemma 3. For each $j \in [d]$, we write

$$\big\{0, x_j^{(1)}, \ldots, x_j^{(n)}, 1\big\} = \big\{v_0^{(j)}, \ldots, v_{n_j}^{(j)}\big\}$$

where $0 = v_0^{(j)} < \cdots < v_{n_j}^{(j)} = 1$. As there can be duplications among $x_j^{(i)}$, n_j may be smaller than $n + 1$.

We construct $f_{\mathbf{c},\boldsymbol{\nu}} \in \mathcal{F}_{\text{tc}}^{d,s}$ (recall (4.1)) that agrees with f at the design points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ explicitly as follows. First, for each $S \subseteq [d]$ with $0 < |S| \leq s$, $\mathbf{l} = (l_j, j \in S) \in \prod_{j \in S}\{0, 1, \ldots, n_j - 1\}$, and $j \in [d]$, we let

$$p_j(S, \mathbf{l}) = \begin{cases} 0 & \text{if } j \notin S \text{ or } l_j = 0 \\ l_j - 1 & \text{otherwise} \end{cases} \quad \text{and} \quad q_j(S, \mathbf{l}) = \begin{cases} 0 & \text{if } j \notin S \\ l_j + 1 & \text{otherwise.} \end{cases}$$

Observe that $0 \leq q_j(S, \mathbf{l}) - p_j(S, \mathbf{l}) \leq 2$ and that $q_j(S, \mathbf{l}) - p_j(S, \mathbf{l}) = 2$ if $j \in S$ and $l_j \neq 0$. It thus follows that $\max_j(q_j(S, \mathbf{l}) - p_j(S, \mathbf{l})) = 2$, unless $\mathbf{l} = \mathbf{0}$. Next, we let $c_0 = f(0, \ldots, 0)$

and, for each S, let

$$c_S = \begin{bmatrix} v^{(1)}_{p_1(S,\mathbf{0})}, \dots, v^{(1)}_{q_1(S,\mathbf{0})} & \\ \vdots & ;\ f \\ v^{(d)}_{p_d(S,\mathbf{0})}, \dots, v^{(d)}_{q_d(S,\mathbf{0})} & \end{bmatrix}. \tag{D.9}$$

Also, for each S, we let ν_S be the discrete measure supported on

$$\Big(\prod_{j\in S}\big\{v^{(j)}_0, \dots, v^{(j)}_{n_j-1}\big\}\Big) \setminus \{\mathbf{0}\}$$

for which

$$\nu_S\big(\big\{\big(v^{(j)}_{l_j}, j\in S\big)\big\}\big) = -\Big(\prod_{\substack{j\in S\\ l_j\neq 0}}\big(v^{(j)}_{l_j+1} - v^{(j)}_{l_j-1}\big)\Big)\begin{bmatrix} v^{(1)}_{p_1(S,\mathbf{l})}, \dots, v^{(1)}_{q_1(S,\mathbf{l})} & \\ \vdots & ;\ f \\ v^{(d)}_{p_d(S,\mathbf{l})}, \dots, v^{(d)}_{q_d(S,\mathbf{l})} & \end{bmatrix} \tag{D.10}$$

for $\mathbf{l} = (l_j, j\in S) \in \prod_{j\in S}\{0,1,\dots,n_j-1\}\setminus\{\mathbf{0}\}$. Since $f\in\mathcal{F}^d_{\mathrm{tcp}}$, we have

$$\begin{bmatrix} v^{(1)}_{p_1(S,\mathbf{l})}, \dots, v^{(1)}_{q_1(S,\mathbf{l})} & \\ \vdots & ;\ f \\ v^{(d)}_{p_d(S,\mathbf{l})}, \dots, v^{(d)}_{q_d(S,\mathbf{l})} & \end{bmatrix} \le 0$$

for every $\mathbf{l}\neq\mathbf{0}$, and this ensures that ν_S is a (nonnegative) measure for each S. Now, we prove that $f_{\mathbf{c},\boldsymbol{\nu}} \in \mathcal{F}^{d,s}_{\mathrm{tc}}$, constructed from these c_0, c_S, and ν_S, agrees with f at the design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. In fact, we can prove a strong statement. We can show that $f_{\mathbf{c},\boldsymbol{\nu}}(v^{(1)}_{i_1}, \dots, v^{(d)}_{i_d}) = f(v^{(1)}_{i_1}, \dots, v^{(d)}_{i_d})$ for all $i_j\in\{0,\dots,n_j\}$, $j\in[d]$. For this, we use the following lemma, which will be proved right after this proof. Here, for each S and $w_j\in[0,1]$ for $j\notin S$, we denote by $f_S(\cdot;(w_j, j\notin S))$ the restriction of f to the subset $x_j = w_j$, $j\notin S$ of $[0,1]^d$. Also, recall the definition of $\triangle$ from (D.2).

Lemma 29. *Suppose we are given $S\subseteq[d]$ and $w_j\in[0,1]$ for each $j\notin S$. Then, we have*

$$\triangle f_S\big(\cdot;(w_j, j\notin S)\big)\Big(\prod_{j\in S}\big[0, v^{(j)}_{i_j}\big]\Big) \tag{D.11}$$

$$= \sum_{\mathbf{l}\in\prod_{j\in S}\{0,\dots,n_j-1\}} \prod_{j\in S}\big(v^{(j)}_{i_j} - v^{(j)}_{l_j}\big)_+ \cdot \prod_{\substack{j\in S\\ l_j\neq 0}}\big(v^{(j)}_{l_j+1} - v^{(j)}_{l_j-1}\big) \cdot \begin{bmatrix} v^{(j)}_{p_j(S,\mathbf{l})}, \dots, v^{(j)}_{q_j(S,\mathbf{l})}, & j\in S & ;\ f \\ w_j, & j\notin S & \end{bmatrix}$$

for every $i_j\in\{0,1,\dots,n_j\}$, $j\in[d]$.

It follows from Lemma 29 that, for each S,

$$\begin{aligned}
&c_S \prod_{j \in S} v_{i_j}^{(j)} - \sum_{\mathbf{l} \in \prod_{j \in S}\{0,\dots,n_j-1\} \setminus \{\mathbf{0}\}} \left(\prod_{j \in S} \big(v_{i_j}^{(j)} - v_{l_j}^{(j)}\big)_+ \right) \cdot \nu_S\big(\big\{\big(v_{l_j}^{(j)}, j \in S\big)\big\}\big) \\
&\quad = \sum_{\mathbf{l} \in \prod_{j \in S}\{0,\dots,n_j-1\}} \left(\prod_{j \in S} \big(v_{i_j}^{(j)} - v_{l_j}^{(j)}\big)_+ \right) \left(\prod_{\substack{j \in S \\ l_j \neq 0}} \big(v_{l_j+1}^{(j)} - v_{l_j-1}^{(j)}\big) \right) \\
&\qquad\qquad \cdot \begin{bmatrix} \begin{matrix} v_{p_j(S,\mathbf{l})}^{(j)}, \dots, v_{q_j(S,\mathbf{l})}^{(j)}, & j \in S \\ 0, & j \notin S \end{matrix} & ; & f \end{bmatrix} \\
&\quad = \triangle f_S\big(\cdot; (0, j \notin S)\big)\left(\prod_{j \in S} \big[0, v_{i_j}^{(j)}\big] \right) = \triangle_S f\left(\prod_{j \in S} \big[0, v_{i_j}^{(j)}\big] \right)
\end{aligned}$$

for every $i_j \in \{0, 1, \dots, n_j\}$, $j \in [d]$. Together with the interaction restriction condition (4.16) for $S \subseteq [d]$ with $|S| > s$, this implies that

$$\begin{aligned}
f_{\mathbf{c},\boldsymbol{\nu}}\big(v_{i_1}^{(1)}, \dots, v_{i_d}^{(d)}\big) &= f(0, \dots, 0) + \sum_{0 < |S| \leq s} c_S \prod_{j \in S} v_{i_j}^{(j)} \\
&\quad - \sum_{0 < |S| \leq s} \sum_{\mathbf{l} \in \prod_{j \in S}\{0,\dots,n_j-1\} \setminus \{\mathbf{0}\}} \left(\prod_{j \in S} \big(v_{i_j}^{(j)} - v_{l_j}^{(j)}\big)_+ \right) \cdot \nu_S\big(\big\{\big(v_{l_j}^{(j)}, j \in S\big)\big\}\big) \\
&= f(0, \dots, 0) + \sum_{0 < |S| \leq s} \triangle_S f\left(\prod_{j \in S} \big[0, v_{i_j}^{(j)}\big] \right) \\
&= f(0, \dots, 0) + \sum_{\emptyset \neq S \subseteq [d]} \triangle_S f\left(\prod_{j \in S} \big[0, v_{i_j}^{(j)}\big] \right) = f\big(v_{i_1}^{(1)}, \dots, v_{i_d}^{(d)}\big)
\end{aligned}$$

for every $i_j \in \{0, 1, \dots, n_j\}$, $j \in [d]$. $\square$

Proof of Lemma 29. We prove the lemma by induction on $|S|$. We first consider the case $|S| = 1$. In this case, the claim of the lemma can be readily checked as follows. Without loss of generality, here, we assume that $S = \{1\}$:

$$\begin{aligned}
\text{(D.11)} &= v_{i_1}^{(1)} \begin{bmatrix} \begin{matrix} v_0^{(1)}, v_1^{(1)} \\ w_2 \\ \vdots \\ w_d \end{matrix} & ; & f \end{bmatrix} + \sum_{l_1=1}^{i_1-1} \big(v_{i_1}^{(1)} - v_{l_1}^{(1)}\big)\big(v_{l_1+1}^{(1)} - v_{l_1-1}^{(1)}\big) \begin{bmatrix} \begin{matrix} v_{l_1-1}^{(1)}, v_{l_1}^{(1)}, v_{l_1+1}^{(1)} \\ w_2 \\ \vdots \\ w_d \end{matrix} & ; & f \end{bmatrix} \\
&= v_{i_1}^{(1)} \begin{bmatrix} \begin{matrix} v_0^{(1)}, v_1^{(1)} \\ w_2 \\ \vdots \\ w_d \end{matrix} & ; & f \end{bmatrix} + \sum_{l_1=1}^{i_1-1} \big(v_{i_1}^{(1)} - v_{l_1}^{(1)}\big) \left[\begin{bmatrix} \begin{matrix} v_{l_1}^{(1)}, v_{l_1+1}^{(1)} \\ w_2 \\ \vdots \\ w_d \end{matrix} & ; & f \end{bmatrix} - \begin{bmatrix} \begin{matrix} v_{l_1-1}^{(1)}, v_{l_1}^{(1)} \\ w_2 \\ \vdots \\ w_d \end{matrix} & ; & f \end{bmatrix} \right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l_1=1}^{i_1} \big(v^{(1)}_{l_1} - v^{(1)}_{l_1-1}\big) \begin{bmatrix} v^{(1)}_{l_1-1}, v^{(1)}_{l_1} & \\ w_2 & \\ \vdots & ;\; f \\ w_d & \end{bmatrix} \\
&= f\big(v^{(1)}_{i_1}, w_2, \dots, w_d\big) - f(0, w_2, \dots, w_d) = \triangle f_{\{1\}}\big(\cdot; (w_2, \dots, w_d)\big)\big(\big[0, v^{(1)}_{i_1}\big]\big).
\end{aligned}
$$

Next, we suppose that the claim holds for all $S \subseteq [d]$ with $|S| = m - 1$ and prove the claim for subsets S with $|S| = m$. For notational convenience, here, we only consider the case $S = [m]$, but the same argument applies to every subset S with $|S| = m$. Through the same computation as in the case $S = \{1\}$, we can show that

$$
\begin{aligned}
\text{(D.11)} &= \sum_{l_1=0}^{n_1-1} \cdots \sum_{l_{m-1}=0}^{n_{m-1}-1} \prod_{j=1}^{m-1} \big(v^{(j)}_{i_j} - v^{(j)}_{l_j}\big)_+ \cdot \prod_{\substack{j \in [m-1] \\ l_j \neq 0}} \big(v^{(j)}_{l_j+1} - v^{(j)}_{l_j-1}\big) \\
&\quad \cdot \Bigg[v^{(m)}_{i_m} \begin{bmatrix} v^{(1)}_{p_1(S,\mathbf{l})}, \dots, v^{(1)}_{q_1(S,\mathbf{l})} & \\ \vdots & \\ v^{(m-1)}_{p_{m-1}(S,\mathbf{l})}, \dots, v^{(m-1)}_{q_{m-1}(S,\mathbf{l})} & \\ v^{(m)}_0, v^{(m)}_1 & ;\; f \\ w_{m+1} & \\ \vdots & \\ w_d & \end{bmatrix} \\
&\quad + \sum_{l_m=1}^{i_m-1} \big(v^{(m)}_{i_m} - v^{(m)}_{l_m}\big)\big(v^{(m)}_{l_m+1} - v^{(m)}_{l_m-1}\big) \begin{bmatrix} v^{(1)}_{p_1(S,\mathbf{l})}, \dots, v^{(1)}_{q_1(S,\mathbf{l})} & \\ \vdots & \\ v^{(m-1)}_{p_{m-1}(S,\mathbf{l})}, \dots, v^{(m-1)}_{q_{m-1}(S,\mathbf{l})} & \\ v^{(m)}_{l_m-1}, v^{(m)}_{l_m}, v^{(m)}_{l_m+1} & ;\; f \\ w_{m+1} & \\ \vdots & \\ w_d & \end{bmatrix} \Bigg] \\
&= \sum_{l_1=0}^{n_1-1} \cdots \sum_{l_{m-1}=0}^{n_{m-1}-1} \prod_{j=1}^{m-1} \big(v^{(j)}_{i_j} - v^{(j)}_{l_j}\big)_+ \cdot \prod_{\substack{j \in [m-1] \\ l_j \neq 0}} \big(v^{(j)}_{l_j+1} - v^{(j)}_{l_j-1}\big)
\end{aligned}
$$

$$
\cdot \left[\left[\begin{matrix} v^{(1)}_{p_1(S,\mathbf{l})}, \dots, v^{(1)}_{q_1(S,\mathbf{l})} \\ \vdots \\ v^{(m-1)}_{p_{m-1}(S,\mathbf{l})}, \dots, v^{(m-1)}_{q_{m-1}(S,\mathbf{l})} \\ v^{(m)}_{i_m} \\ w_{m+1} \\ \vdots \\ w_d \end{matrix} \; ; \; f \right] - \left[\begin{matrix} v^{(1)}_{p_1(S,\mathbf{l})}, \dots, v^{(1)}_{q_1(S,\mathbf{l})} \\ \vdots \\ v^{(m-1)}_{p_{m-1}(S,\mathbf{l})}, \dots, v^{(m-1)}_{q_{m-1}(S,\mathbf{l})} \\ 0 \\ w_{m+1} \\ \vdots \\ w_d \end{matrix} \; ; \; f \right] \right]
$$

$$
\begin{aligned}
&= \triangle f_{[m-1]}\big(\cdot; \big(v^{(m)}_{i_m}, w_{m+1}, \dots, w_d\big)\big) \left(\prod_{j=1}^{m-1} \big[0, v^{(j)}_{i_j}\big] \right) \\
&\quad - \triangle f_{[m-1]}(\cdot; (0, w_{m+1}, \dots, w_d)) \left(\prod_{j=1}^{m-1} \big[0, v^{(j)}_{i_j}\big] \right) \\
&= \triangle f_{[m]}\big(\cdot; (w_{m+1}, \dots, w_d)\big) \left(\prod_{j=1}^{m} \big[0, v^{(j)}_{i_j}\big] \right),
\end{aligned}
$$

where the second-to-last equality is due to the induction hypothesis. This proves the claim of the lemma for $S = [m]$ and concludes the proof. □

D.3 Proof of Proposition 7

Proposition 7 can be proved using Lemma 10 in Appendix C, which is a simple consequence of the fundamental theorem of calculus.

Proof of Proposition 7. Suppose $f^{(\mathbf{p})}$ exists and is continuous on $[0,1]^d$ for every $\mathbf{p} \in \{0,1,2\}^d$. Assume further that $f^{(\mathbf{p})} = 0$ for all $\mathbf{p} \in \{0,1\}^d$ with $\sum_j p_j > s$. For each nonempty $S \subseteq [d]$, let $\mathbf{p}_S$ denote the d-dimensional vector for which $(p_S)_j = 1$ if $j \in S$, and $(p_S)_j = 0$ otherwise. Also, for each S, let f_S be the restriction of $f^{(\mathbf{p}_S)}$ to $\prod_{j \in S}[0,1] \times \prod_{j \notin S}\{0\}$. Note that $f_S = 0$ for every S with $|S| > s$. Since f is smooth in the sense of Lemma 10, we can write it as

$$
\begin{aligned}
f(x_1, \dots, x_d) &= f(0, \dots, 0) + \sum_{\emptyset \neq S \subseteq [d]} \int_{\prod_{j \in S}[0, x_j]} f_S(t_j, j \in S) \, d(t_j, j \in S) \\
&= f(0, \dots, 0) + \sum_{0 < |S| \leq s} \int_{\prod_{j \in S}[0, x_j]} f_S(t_j, j \in S) \, d(t_j, j \in S)
\end{aligned}
$$

for $(x_1, \dots, x_d) \in [0,1]^d$. Also, since f_S are continuous, if $-f_S$ are entirely monotone, then we can apply Lemma 25 to conclude that $f \in \mathcal{F}^{d,s}_{\mathrm{tc}}$.

We now show that $-f_S$ is entirely monotone for every S with $0 < |S| \leq s$. Without loss of generality, we verify this only for $S = [m]$. Since f_S is smooth in the sense of Lemma 10, it can be expressed as

$$f_S(x_1, \dots, x_m) = f_S(0, \dots, 0) + \sum_{\mathbf{p} \in \{0,1\}^m \setminus \{\mathbf{0}\}} \int_{\prod_{j \in S_{\mathbf{p}}} [0, x_j]} f_S^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})}) \, d\mathbf{t}^{(\mathbf{p})}$$

for $(x_1, \dots, x_m) \in [0,1]^m$. For each $\mathbf{l} = (l_1, \dots, l_m) \in \{0,1\}^m \setminus \{\mathbf{0}\}$, the divided difference of f_S of order $\mathbf{l}$ is thus given by

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{l_1+1}^{(1)} & \\ \vdots & ; \ f_S \\ v_1^{(m)}, \dots, v_{l_m+1}^{(m)} & \end{bmatrix} = \sum_{\mathbf{p} : S_{\mathbf{l}} \subseteq S_{\mathbf{p}}} \int_{\prod_{j \in S_{\mathbf{l}}} (v_1^{(j)}, v_2^{(j)}] \times \prod_{j \in S_{\mathbf{p}} \setminus S_{\mathbf{l}}} [0, v_1^{(j)}]} f_S^{(\mathbf{p})}(\widetilde{\mathbf{t}}^{(\mathbf{p})}) \, d\mathbf{t}^{(\mathbf{p})}.$$

Note that $f_S^{(\mathbf{p})}$ is the restriction of $f^{(p_1+1, \dots, p_m+1, 0, \dots, 0)}$ to $[0,1]^m \times \{0\}^{d-m}$ and that $\max_j (p_j + 1) = 2$. Hence, by the assumption that $f^{(\mathbf{p})} \leq 0$ for every $\mathbf{p} \in \{0,1,2\}^d$ with $\max_j p_j = 2$, $f_S^{(\mathbf{p})} \leq 0$ for every $\mathbf{p} \in \{0,1\}^m \setminus \{\mathbf{0}\}$. This ensures that the divided difference of f_S of order $\mathbf{l}$ is nonpositive for every $\mathbf{l} \in \{0,1\}^m \setminus \{\mathbf{0}\}$, which means that $-f_S$ is entirely monotone. □

D.4 Proof of Lemma 4

Proof of Lemma 4. In this proof, we reuse $f_{\mathbf{c},\boldsymbol{\nu}}$ that we constructed in our proof of Lemma 3. Recall our choices of c_0, c_S, and ν_S (with $s = d$) from (D.9) and (D.10). Note that, since

$$\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} = \prod_{j=1}^{d} \Big\{0, \frac{1}{n_j}, \dots, \frac{n_j - 1}{n_j}\Big\},$$

here, we have

$$\{v_0^{(j)}, \dots, v_{n_j}^{(j)}\} = \Big\{0, \frac{1}{n_j}, \dots, \frac{n_j - 1}{n_j}, 1\Big\}$$

for each $j \in [d]$. Since $f_{\mathbf{c},\boldsymbol{\nu}}$ and f have the same values at all design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, the definition of $V_{\mathrm{design}}(\cdot)$ implies that

$$V_{\mathrm{design}}(f) \leq V_{\mathrm{mars}}(f_{\mathbf{c},\boldsymbol{\nu}}) = \sum_{\emptyset \neq S \subseteq [d]} \nu_S\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big).$$

Hence, it suffices to show that

$$\nu_S\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big) = \begin{bmatrix} 0, 1/n_j, & j \in S & ; \ f \\ 0, & j \notin S & \end{bmatrix} - \begin{bmatrix} (n_j - 1)/n_j, 1, & j \in S & ; \ f \\ 0, & j \notin S & \end{bmatrix} \tag{D.12}$$

for each nonempty subset $S \subseteq [d]$. Here, we prove this equation only for $S = [m]$ for notational convenience, but the other cases can also be proved exactly the same. When $S = [m]$, our choice (D.10) of ν_S leads to

$$\nu_S\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big) - \begin{bmatrix} 0, 1/n_j, & j \in S & ; \ f \\ 0, & j \notin S & \end{bmatrix}$$

$$= -\sum_{l_1=0}^{n_1-1} \cdots \sum_{l_m=0}^{n_m-1} \Bigg(\prod_{\substack{j\in[m] \\ l_j\neq 0}} \frac{2}{n_j} \Bigg) \begin{bmatrix} p_1(S,\mathbf{l})/n_1, \dots, q_1(S,\mathbf{l})/n_1 & \\ \vdots & \\ p_m(S,\mathbf{l})/n_m, \dots, q_m(S,\mathbf{l})/n_m & ; \;\; f \\ 0 & \\ \vdots & \\ 0 & \end{bmatrix}$$

$$= -\sum_{l_1=0}^{n_1-1} \cdots \sum_{l_{m-1}=0}^{n_{m-1}-1} \Bigg(\prod_{\substack{j\in[m-1] \\ l_j\neq 0}} \frac{2}{n_j} \Bigg) \Bigg[\begin{bmatrix} p_1(S,\mathbf{l})/n_1, \dots, q_1(S,\mathbf{l})/n_1 & \\ \vdots & \\ p_{m-1}(S,\mathbf{l})/n_{m-1}, \dots, q_{m-1}(S,\mathbf{l})/n_{m-1} & \\ 0, 1/n_m & ; \;\; f \\ 0 & \\ \vdots & \\ 0 & \end{bmatrix}$$

$$+ \sum_{l_m=1}^{n_m-1} \frac{2}{n_m} \begin{bmatrix} p_1(S,\mathbf{l})/n_1, \dots, q_1(S,\mathbf{l})/n_1 & \\ \vdots & \\ p_{m-1}(S,\mathbf{l})/n_{m-1}, \dots, q_{m-1}(S,\mathbf{l})/n_{m-1} & \\ (l_m-1)/n_m, l_m/n_m, (l_m+1)/n_m & ; \;\; f \\ 0 & \\ \vdots & \\ 0 & \end{bmatrix} \Bigg]$$

$$= -\sum_{l_1=0}^{n_1-1} \cdots \sum_{l_{m-1}=0}^{n_{m-1}-1} \Bigg(\prod_{\substack{j\in[m-1] \\ l_j\neq 0}} \frac{2}{n_j} \Bigg) \Bigg[\begin{bmatrix} p_1(S,\mathbf{l})/n_1, \dots, q_1(S,\mathbf{l})/n_1 & \\ \vdots & \\ p_{m-1}(S,\mathbf{l})/n_{m-1}, \dots, q_{m-1}(S,\mathbf{l})/n_{m-1} & \\ 0, 1/n_m & ; \;\; f \\ 0 & \\ \vdots & \\ 0 & \end{bmatrix}$$

$$+ \sum_{l_m=1}^{n_m-1} \Bigg(\begin{bmatrix} p_1(S,\mathbf{l})/n_1, \dots, q_1(S,\mathbf{l})/n_1 & \\ \vdots & \\ p_{m-1}(S,\mathbf{l})/n_{m-1}, \dots, q_{m-1}(S,\mathbf{l})/n_{m-1} & \\ l_m/n_m, (l_m+1)/n_m & ; \;\; f \\ 0 & \\ \vdots & \\ 0 & \end{bmatrix}$$

$$
- \left[\begin{array}{cc} p_1(S,\mathbf{l})/n_1, \dots, q_1(S,\mathbf{l})/n_1 & \\ \vdots & \\ p_{m-1}(S,\mathbf{l})/n_{m-1}, \dots, q_{m-1}(S,\mathbf{l})/n_{m-1} & \\ (l_m-1)/n_m, l_m/n_m & ; \ f \\ 0 & \\ \vdots & \\ 0 & \end{array}\right]\Bigg)\Bigg]
$$

$$
= - \sum_{l_1=0}^{n_1-1} \cdots \sum_{l_{m-1}=0}^{n_{m-1}-1} \Bigg(\prod_{\substack{j \in [m-1] \\ l_j \neq 0}} \frac{2}{n_j} \Bigg) \left[\begin{array}{cc} p_1(S,\mathbf{l})/n_1, \dots, q_1(S,\mathbf{l})/n_1 & \\ \vdots & \\ p_{m-1}(S,\mathbf{l})/n_{m-1}, \dots, q_{m-1}(S,\mathbf{l})/n_{m-1} & \\ (n_m-1)/n_m, 1 & ; \ f \\ 0 & \\ \vdots & \\ 0 & \end{array}\right].
$$

Repeating this argument with the $(m-1)^{\text{th}}, \dots, 1^{\text{st}}$ coordinates in turn, we can obtain

$$
\begin{aligned}
&\nu_S\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big) - \left[\begin{array}{ccc} 0, 1/n_j, & j \in S & \\ 0, & j \notin S & ; \ f \end{array}\right] \\
&= - \left[\begin{array}{cc} (n_1-1)/n_1, 1 & \\ \vdots & \\ (n_m-1)/n_m, 1 & ; \ f \\ 0 & \\ \vdots & \\ 0 & \end{array}\right] = - \left[\begin{array}{ccc} (n_j-1)/n_j, 1, & j \in S & \\ 0, & j \notin S & ; \ f \end{array}\right],
\end{aligned}
$$

which proves (D.12) for the case $S = [m]$. □

D.5 Proof of Theorem 9

The following theorem, established in [28], plays a central role in our proof of Theorem 9. Before presenting the theorem, we briefly outline the problem setting considered in [28].

Consider the model

$$
y_i = \theta_i^* + \xi_i \quad \text{for } i = 1, \dots, n, \tag{D.13}
$$

where $\mathbf{y} = (y_1, \dots, y_n)$ is the observation vector, $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_n^*)$ is the unknown vector to be estimated, and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ is a vector of Gaussian errors with $\xi_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Also, let $K \subseteq \mathbb{R}^n$ be a closed convex set. The least squares estimator of $\boldsymbol{\theta}^*$ over K based on the observations $y_1, \dots, y_n$ is then defined as

$$
\hat{\boldsymbol{\theta}}_K = \operatorname*{argmin}_{\boldsymbol{\theta} \in K} \|\mathbf{y} - \boldsymbol{\theta}\|_2, \tag{D.14}
$$

where $\|\cdot\|_2$ denotes the Euclidean norm.

[28] studied the risk of this estimator $\hat{\boldsymbol{\theta}}_K$, measured by the expected squared error $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}^*\|_2^2]$, where the expectation is taken over the Gaussian errors ξ_i. A key tool in their analysis is the function h defined as

$$h(t) = \mathbb{E}\Big[\sup_{\boldsymbol{\theta}\in K:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le t}\langle \boldsymbol{\xi}, \boldsymbol{\theta}-\boldsymbol{\theta}^*\rangle\Big] - \frac{t^2}{2} \tag{D.15}$$

for $t \ge 0$. If no $\boldsymbol{\theta} \in K$ satisfies $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 \le t$, $h(t)$ is set to $-\infty$. [28, Theorem 1] proved that this function h has a unique maximizer, denoted by t^*. Furthermore, [28, Corollary 1.2], restated as Theorem 28 below, provides an upper bound on the expected squared error $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}^*\|_2^2]$ in terms of t^*.

Theorem 28 (Corollary 1.2 of [28])**.** *The expected squared error can be bounded as*

$$\mathbb{E}\big[\|\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}^*\|_2^2\big] \le C(\sigma^2 + t^{*2})$$

for some universal constant C.

However, determining the maximizer t^* of the function h is often quite challenging. Fortunately, we can utilize the following result in [28] to obtain an upper bound for t^*, which can then be used instead to bound the expected squared error.

Proposition 16 (Proposition 1.3 of [28])**.** *If $h(t_1) \ge h(t_2)$ for some $0 \le t_1 < t_2$, then $t^* \le t_2$.*

Remark 15 (Well-specified case)**.** *If $\boldsymbol{\theta}^* \in K$, it holds that $h(0) = 0$. Thus, in this case, for every $t_2 > 0$ with $h(t_2) \le 0$, we have $t^* \le t_2$.*

We now establish a connection between $\hat{f}_n^{d,s}$ and $\hat{\boldsymbol{\theta}}_K$ with an appropriate closed convex set K. Consider the index set

$$I_0 = \prod_{j=1}^{d}\{0, 1, \dots, n_j - 1\}.$$

We first re-index $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ using I_0, so that $\mathbf{x}^{(\mathbf{i})} = (i_1/n_1, \dots, i_d/n_d)$ for $\mathbf{i} = (i_1, \dots, i_d) \in I_0$. Similarly, the observations $y_1, \dots, y_n$ are re-indexed as $(y_{\mathbf{i}}, \mathbf{i} \in I_0)$. Next, define

$$T = \big\{(f(\mathbf{x}^{(\mathbf{i})}), \mathbf{i} \in I_0) : f \in \mathcal{F}_{\mathrm{tc}}^{d,s}\big\}. \tag{D.16}$$

Note that T is a closed convex subset of $\mathbb{R}^{|I_0|}$. Also, let $\hat{\boldsymbol{\theta}}_T \in \mathbb{R}^{|I_0|}$ denote the least squares estimator of $\boldsymbol{\theta}^*$ over T:

$$\hat{\boldsymbol{\theta}}_T = \operatorname*{argmin}_{\boldsymbol{\theta}\in T}\|\mathbf{y} - \boldsymbol{\theta}\|_2,$$

where $\mathbf{y} = (y_{\mathbf{i}}, \mathbf{i} \in I_0)$. Then, by the definition (D.16) of T, it is clear that the values of $\hat{f}_n^{d,s}$ at the design points are equal to the components of $\hat{\boldsymbol{\theta}}_T$. That is,

$$\hat{f}_n^{d,s}\Big(\frac{i_1}{n_1}, \dots, \frac{i_d}{n_d}\Big) = (\hat{\theta}_T)_{\mathbf{i}}$$

for all $\mathbf{i} = (i_1, \dots, i_d) \in I_0$. Consequently, the risk $R_F(\hat{f}_n^{d,s}, f^*)$ of $\hat{f}_n^{d,s}$ can be expressed in terms of the expected squared error of $\hat{\boldsymbol{\theta}}_T$:

$$R_F(\hat{f}_n^{d,s}, f^*) = \frac{1}{n} \cdot \mathbb{E}\big[\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2^2\big],$$

where $\boldsymbol{\theta}^* = (f^*(\mathbf{x}^{(\mathbf{i})}), \mathbf{i} \in I_0)$. This relationship provides a way to analyze the risk $R_F(\hat{f}_n^{d,s}, f^*)$ of $\hat{f}_n^{d,s}$ by studying the expected squared error $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2^2]$ of $\hat{\boldsymbol{\theta}}_T$.

We also employ the following concepts and notations in our proof of Theorem 9. For $\boldsymbol{\theta} \in \mathbb{R}^{|I_0|}$ and a vector of nonnegative integers $\mathbf{p} = (p_1, \dots, p_d)$, the discrete difference of $\boldsymbol{\theta}$ of order $\mathbf{p}$ at $(i_1, \dots, i_d)$, where $(i_1, \dots, i_d) \in I_0$ and $i_j \geq p_j$ for $j \in [d]$, is defined as

$$(D^{(\mathbf{p})}\theta)_{i_1,\dots,i_d} = \frac{\prod_{j=1}^d p_j!}{\prod_{j=1}^d n_j^{p_j}} \cdot \begin{bmatrix} (i_1 - p_1)/n_1, \dots, i_1/n_1 & \\ \vdots & ; \; f \\ (i_d - p_d)/n_d, \dots, i_d/n_d & \end{bmatrix}, \tag{D.17}$$

where $f : [0,1]^d \to \mathbb{R}$ is any function satisfying $f(l_1/n_1, \dots, l_d/n_d) = \theta_{\mathbf{l}}$ for all $\mathbf{l} = (l_1, \dots, l_d) \in I_0$. Since (D.17) only depends on the values of f at $(l_1/n_1, \dots, l_d/n_d)$ for $\mathbf{l} = (l_1, \dots, l_d) \in I_0$, the discrete difference $(D^{(\mathbf{p})}\theta)_{i_1,\dots,i_d}$ is well-defined; it is independent of the choice of f.

Example 5 (d = 2)**.** *Let* $\boldsymbol{\theta} = (\theta_{\mathbf{i}}, \mathbf{i} \in I_0) \in \mathbb{R}^{|I_0|}$*. For* $\mathbf{p} = (1, 0)$*:*

$$(D^{(1,0)}\theta)_{i_1,i_2} = \frac{1}{n_1}\begin{bmatrix} (i_1 - 1)/n_1, i_1/n_1 & ; \; f \\ i_2/n_2 & \end{bmatrix} = f\Big(\frac{i_1}{n_1}, \frac{i_2}{n_2}\Big) - f\Big(\frac{i_1 - 1}{n_1}, \frac{i_2}{n_2}\Big) = \theta_{i_1,i_2} - \theta_{i_1-1,i_2}$$

provided that $i_1 \geq 1$*. For* $\mathbf{p} = (1, 1)$*:*

$$\begin{aligned}(D^{(1,1)}\theta)_{i_1,i_2} &= \frac{1}{n_1 n_2}\begin{bmatrix} (i_1 - 1)/n_1, i_1/n_1 & ; \; f \\ (i_2 - 1)/n_2, i_2/n_2 & \end{bmatrix} \\ &= f\Big(\frac{i_1}{n_1}, \frac{i_2}{n_2}\Big) - f\Big(\frac{i_1 - 1}{n_1}, \frac{i_2}{n_2}\Big) - f\Big(\frac{i_1}{n_1}, \frac{i_2 - 1}{n_2}\Big) + f\Big(\frac{i_1 - 1}{n_1}, \frac{i_2 - 1}{n_2}\Big) \\ &= \theta_{i_1,i_2} - \theta_{i_1-1,i_2} - \theta_{i_1,i_2-1} + \theta_{i_1-1,i_2-1}\end{aligned}$$

provided that $i_1, i_2 \geq 1$*. For* $\mathbf{p} = (2, 0)$*:*

$$\begin{aligned}(D^{(2,0)}\theta)_{i_1,i_2} &= \frac{2}{n_1^2}\begin{bmatrix} (i_1 - 2)/n_1, (i_1 - 1)/n_1, i_1/n_1 & ; \; f \\ i_2/n_2 & \end{bmatrix} \\ &= f\Big(\frac{i_1}{n_1}, \frac{i_2}{n_2}\Big) - 2f\Big(\frac{i_1 - 1}{n_1}, \frac{i_2}{n_2}\Big) + f\Big(\frac{i_1 - 2}{n_1}, \frac{i_2}{n_2}\Big) = \theta_{i_1,i_2} - 2\theta_{i_1-1,i_2} + \theta_{i_1-2,i_2}\end{aligned}$$

provided that $i_1 \geq 2$*.*

If $\boldsymbol{\theta} \in T$, then we can select f satisfying the required condition from $\mathcal{F}_{\mathrm{tc}}^{d,s}$. Let f_0 denote such a function. Since the divided difference of f_0 of order $\mathbf{p}$ is nonpositive for every $\mathbf{p} \in \{0,1,2\}^d$ with $\max_j p_j = 2$, it then follows that $(D^{(\mathbf{p})}\theta)_{\mathbf{i}} \leq 0$ for all $\mathbf{p} \in \{0,1,2\}^d$ with $\max_j p_j = 2$ and $\mathbf{i} = (i_1, \dots, i_d) \in I_0$ with $i_j \geq p_j$ for $j \in [d]$. Furthermore, using the interaction restriction condition (4.16) that f_0 satisfies for every subset $S \subseteq [d]$ with $|S| > s$, we can show that $(D^{(\mathbf{p})}\theta)_{\mathbf{i}} = 0$ for every $\mathbf{p} \in \{0,1\}^d$ with $\sum_{j=1}^d p_j > s$ and $\mathbf{i} = (i_1, \dots, i_d) \in I_0$ with $i_j \geq p_j$ for $j \in [d]$.

Additionally, for a vector of positive integers $\mathbf{m} = (m_1, \dots, m_d)$, define

$$I(\mathbf{m}) = \prod_{j=1}^d \{0, 1, \dots, m_j - 1\}. \tag{D.18}$$

Note that $I_0 = I(n_1, \dots, n_d)$. For $\boldsymbol{\theta} \in \mathbb{R}^{|I(\mathbf{m})|}$ and a vector of nonnegative integers $\mathbf{p} = (p_1, \dots, p_d)$, the discrete difference of $\boldsymbol{\theta}$ of order $\mathbf{p}$ at $(i_1, \dots, i_d)$, where $(i_1, \dots, i_d) \in I(\mathbf{m})$ and $i_j \geq p_j$ for $j \in [d]$, is defined analogously to (D.17).

Proof of Theorem 9. Let h be the function defined in (D.15) with $K = T$. We aim to find $t_2 > 0$ such that $h(t_2) \leq 0$. After that, we will apply Theorem 28 and Proposition 16 (with $K = T$) to derive an upper bound for the expected squared error $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2^2]$.

To focus on the first term of $h(\cdot)$, we define

$$H(t) = h(t) + \frac{t^2}{2} = \mathbb{E}\Big[\sup_{\boldsymbol{\theta} \in T : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq t} \langle \boldsymbol{\xi}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle\Big]$$

for $t \geq 0$. For technical reasons, we decompose the index set I_0 into 2^d subsets with approximately equal size. For each $j \in [d]$, define

$$J_0^{(j)} = \Big\{i_j \in \mathbb{Z} : 0 \leq i_j \leq \frac{n_j}{2} - 1\Big\} \text{ and } J_1^{(j)} = \Big\{i_j \in \mathbb{Z} : \frac{n_j - 1}{2} \leq i_j \leq n_j - 1\Big\}.$$

Also, for each $\boldsymbol{\delta} \in \{0,1\}^d$, let

$$J_{\boldsymbol{\delta}} = \prod_{j=1}^d J_{\delta_j}^{(j)}.$$

It is clear that $I_0 = \biguplus_{\boldsymbol{\delta} \in \{0,1\}^d} J_{\boldsymbol{\delta}}$, where $\biguplus$ denotes the disjoint union. It thus follows that

$$\begin{aligned} H(t) &= \mathbb{E}\Big[\sup_{\boldsymbol{\theta} \in T : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq t} \sum_{\boldsymbol{\delta} \in \{0,1\}^d} \sum_{\mathbf{i} \in J_{\boldsymbol{\delta}}} \xi_{\mathbf{i}}(\theta_{\mathbf{i}} - \theta_{\mathbf{i}}^*)\Big] \\ &\leq \sum_{\boldsymbol{\delta} \in \{0,1\}^d} \mathbb{E}\Big[\sup_{\boldsymbol{\theta} \in T : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq t} \sum_{\mathbf{i} \in J_{\boldsymbol{\delta}}} \xi_{\mathbf{i}}(\theta_{\mathbf{i}} - \theta_{\mathbf{i}}^*)\Big] = \sum_{\boldsymbol{\delta} \in \{0,1\}^d} H_{\boldsymbol{\delta}}(t), \end{aligned}$$

where

$$H_{\boldsymbol{\delta}}(t) = \mathbb{E}\bigg[\sup_{\boldsymbol{\theta}\in T:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le t}\sum_{\mathbf{i}\in J_{\boldsymbol{\delta}}}\xi_{\mathbf{i}}(\theta_{\mathbf{i}}-\theta^*_{\mathbf{i}})\bigg] \quad \text{for each } \boldsymbol{\delta}\in\{0,1\}^d.$$

Here, we focus on bounding $H_{\mathbf{0}}(\cdot)$, as the bounds for the other $H_{\boldsymbol{\delta}}(\cdot)$ terms can be obtained analogously. We further split the index set $J_{\mathbf{0}}$ into dyadic pieces. For each $j\in[d]$, let R_j denote the largest integer r_j for which the set

$$L^{(j)}_{r_j} := \Big\{i_j\in\mathbb{Z} : \frac{n_j}{2^{r_j+1}}-1 < i_j \le \frac{n_j}{2^{r_j}}-1\Big\} := \big\{a^{(j)}_{r_j}, a^{(j)}_{r_j}+1,\dots,b^{(j)}_{r_j}\big\}$$

is nonempty. Observe that $L^{(j)}_{R_j} = \{0\}$ and $2^{R_j}\le n_j < 2^{R_j+1}$ for all $j\in[d]$. Define $\mathcal{R} = \prod_{j=1}^d\{1,\dots,R_j\}$. Then, $J_{\mathbf{0}}$ can be decomposed as follows:

$$J_{\mathbf{0}} = \biguplus_{\mathbf{r}\in\mathcal{R}} L_{\mathbf{r}}, \quad \text{where} \quad L_{\mathbf{r}} := \prod_{j=1}^d L^{(j)}_{r_j}.$$

For $\boldsymbol{\theta} = (\theta_{\mathbf{i}}, \mathbf{i}\in I_0)\in\mathbb{R}^{|I_0|}$ and $\mathbf{r} = (r_1,\dots,r_d)\in\prod_{j=1}^d\{0,1,\dots,R_j\}$, let $\boldsymbol{\theta}^{(\mathbf{r})}$ denote the subvector of $\boldsymbol{\theta}$ consisting of the components whose indices belong to $L_{\mathbf{r}}$.

Now, define

$$W = \bigg\{(w_{\mathbf{r}},\mathbf{r}\in\mathcal{R}) : w_{\mathbf{r}}\in\{1,\dots,|\mathcal{R}|\} \text{ for each } \mathbf{r}\in\mathcal{R} \text{ and } \sum_{\mathbf{r}\in\mathcal{R}} w_{\mathbf{r}}\le 2|\mathcal{R}|\bigg\}.$$

Also, for each $\mathbf{w} = (w_{\mathbf{r}},\mathbf{r}\in\mathcal{R})\in W$, define

$$\Lambda_{\mathbf{w}}(t) = \Big\{\boldsymbol{\theta}\in T : \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le t \text{ and } \|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2^2 \le \frac{w_{\mathbf{r}}t^2}{|\mathcal{R}|} \text{ for each } \mathbf{r}\in\mathcal{R}\Big\}.$$

We then claim that

$$\{\boldsymbol{\theta}\in T : \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le t\} \subseteq \bigcup_{\mathbf{w}\in W}\Lambda_{\mathbf{w}}(t). \tag{D.19}$$

This can be easily verified as follows. Fix $\boldsymbol{\theta}\in T$ and assume that $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le t$. For each $\mathbf{r}\in\mathcal{R}$, note that

$$\|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2 \le \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le t.$$

Thus, there exists $w_{\mathbf{r}}\in\{1,\dots,|\mathcal{R}|\}$ such that

$$\frac{(w_{\mathbf{r}}-1)t^2}{|\mathcal{R}|} \le \|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2^2 \le \frac{w_{\mathbf{r}}t^2}{|\mathcal{R}|}.$$

Furthermore, since

$$t^2 \ge \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \ge \sum_{\mathbf{r}\in\mathcal{R}}\|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2^2 \ge \frac{t^2}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}}(w_{\mathbf{r}}-1) = t^2\bigg(\frac{1}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}}w_{\mathbf{r}}-1\bigg),$$

it follows that $\sum_{\mathbf{r}\in\mathcal{R}} w_{\mathbf{r}} \leq 2|\mathcal{R}|$. Consequently, $\boldsymbol{\theta} \in \Lambda_{\mathbf{w}}(t)$, completing the verification.

By (D.19), we have

$$H_{\mathbf{0}}(t) \leq \mathbb{E}\bigg[\max_{\mathbf{w}\in W} \sup_{\boldsymbol{\theta}\in\Lambda_{\mathbf{w}}(t)} \sum_{\mathbf{i}\in J_{\mathbf{0}}} \xi_{\mathbf{i}}(\theta_{\mathbf{i}} - \theta^*_{\mathbf{i}})\bigg].$$

Using the following lemma, which is based on the concentration inequality for Gaussian random variables, we can switch the order of the expectation and maximum (at the cost of additional terms):

$$H_{\mathbf{0}}(t) \leq \max_{\mathbf{w}\in W} \mathbb{E}\bigg[\sup_{\boldsymbol{\theta}\in\Lambda_{\mathbf{w}}(t)} \sum_{\mathbf{i}\in J_{\mathbf{0}}} \xi_{\mathbf{i}}(\theta_{\mathbf{i}} - \theta^*_{\mathbf{i}})\bigg] + \sigma t\sqrt{2\log|W|} + \sigma t\sqrt{\frac{\pi}{2}}.$$

Lemma 30 (Lemma D.1 of [68])**.** *Let $\Theta_1, \ldots, \Theta_m$ be subsets of $\mathbb{R}^n$ containing the origin. Then, we have*

$$\mathbb{E}\Big[\max_{i=1,\ldots,m} \sup_{\boldsymbol{\theta}\in\Theta_i} \langle \boldsymbol{\xi}, \boldsymbol{\theta}\rangle\Big] \leq \max_{i=1,\ldots,m} \mathbb{E}\Big[\sup_{\boldsymbol{\theta}\in\Theta_i} \langle \boldsymbol{\xi}, \boldsymbol{\theta}\rangle\Big] + \sigma\Big(\sqrt{2\log m} + \sqrt{\frac{\pi}{2}}\Big)\cdot\Big(\max_{i=1,\ldots,m} \sup_{\boldsymbol{\theta}\in\Theta_i} \|\boldsymbol{\theta}\|_2\Big),$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ is a vector of Gaussian errors with $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

Since

$$|W| \leq \sum_{k=|\mathcal{R}|}^{2|\mathcal{R}|} \binom{k-1}{k-|\mathcal{R}|} \leq \sum_{k=|\mathcal{R}|}^{2|\mathcal{R}|} \binom{2|\mathcal{R}|-1}{k-|\mathcal{R}|} \leq 2^{2|\mathcal{R}|-1},$$

it thus follows that

$$H_{\mathbf{0}}(t) \leq \max_{\mathbf{w}\in W} \mathbb{E}\bigg[\sup_{\boldsymbol{\theta}\in\Lambda_{\mathbf{w}}(t)} \sum_{\mathbf{i}\in J_{\mathbf{0}}} \xi_{\mathbf{i}}(\theta_{\mathbf{i}} - \theta^*_{\mathbf{i}})\bigg] + 2\sigma t\sqrt{|\mathcal{R}|} + \sigma t\sqrt{\frac{\pi}{2}}. \tag{D.20}$$

Fix $\mathbf{w} = (w_{\mathbf{r}}, \mathbf{r} \in \mathcal{R}) \in W$. Since $J_{\mathbf{0}} = \biguplus_{\mathbf{r}\in\mathcal{R}} L_{\mathbf{r}}$, we have

$$\begin{aligned}
\mathbb{E}\bigg[\sup_{\boldsymbol{\theta}\in\Lambda_{\mathbf{w}}(t)} \sum_{\mathbf{i}\in J_{\mathbf{0}}} \xi_{\mathbf{i}}(\theta_{\mathbf{i}} - \theta^*_{\mathbf{i}})\bigg] &= \mathbb{E}\bigg[\sup_{\boldsymbol{\theta}\in\Lambda_{\mathbf{w}}(t)} \sum_{\mathbf{r}\in\mathcal{R}} \langle \boldsymbol{\xi}^{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})} - \boldsymbol{\theta}^{*(\mathbf{r})}\rangle\bigg] \\
&\leq \sum_{\mathbf{r}\in\mathcal{R}} \mathbb{E}\bigg[\sup_{\boldsymbol{\theta}\in\Lambda_{\mathbf{w}}(t)} \langle \boldsymbol{\xi}^{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})} - \boldsymbol{\theta}^{*(\mathbf{r})}\rangle\bigg] \leq \sum_{\mathbf{r}\in\mathcal{R}} \mathbb{E}\bigg[\sup_{\substack{\boldsymbol{\theta}\in T: \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\leq t \\ \|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2 \leq t(w_{\mathbf{r}}/|\mathcal{R}|)^{1/2}}} \langle \boldsymbol{\xi}^{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})} - \boldsymbol{\theta}^{*(\mathbf{r})}\rangle\bigg].
\end{aligned}$$

For each $\mathbf{r} \in \mathcal{R}$, we now bound

$$\mathbb{E}\bigg[\sup_{\substack{\boldsymbol{\theta}\in T: \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\leq t \\ \|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2 \leq t(w_{\mathbf{r}}/|\mathcal{R}|)^{1/2}}} \langle \boldsymbol{\xi}^{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})} - \boldsymbol{\theta}^{*(\mathbf{r})}\rangle\bigg]$$

using the following theorem (Theorem 29) established in the proof of Theorem 5.1 of [78] (an earlier version of [79]).

For a vector of positive integers $\mathbf{m} = (m_1, \ldots, m_d)$ and $\boldsymbol{\theta} = (\theta_{\mathbf{i}}, \mathbf{i} \in I(\mathbf{m}))$, define

$$V(\boldsymbol{\theta}) = \sum_{\substack{\mathbf{p} \in \{0,1,2\}^d \\ \max_j p_j = 2}} \left[\prod_{j=1}^{d} m_j^{p_j - \mathbf{1}(p_j = 2)} \cdot \sum_{\mathbf{i} \in I^{(\mathbf{p})}(\mathbf{m})} |(D^{(\mathbf{p})}\theta)_{\mathbf{i}}| \right],$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. Here, for each $\mathbf{p} \in \{0,1,2\}^d$ with $\max_j p_j = 2$, the index set $I^{(\mathbf{p})}(\mathbf{m})$ is given by $I^{(\mathbf{p})}(\mathbf{m}) = I_1^{(\mathbf{p})} \times \cdots \times I_d^{(\mathbf{p})}$, where

$$I_j^{(\mathbf{p})} = \begin{cases} \{p_j\} & \text{if } p_j \neq 2, \\ \{2, 3, \ldots, m_j - 1\} & \text{if } p_j = 2 \end{cases} \quad \text{for } j \in [d].$$

This $V(\cdot)$ can be viewed as a discrete analogue of $V_{\text{mars}}(\cdot)$ in (4.19). Also, it is straightforward to verify that if $(D^{(\mathbf{p})}\theta)_{\mathbf{i}} \leq 0$ for all $\mathbf{i} \in I^{(\mathbf{p})}(\mathbf{m})$ and $\mathbf{p} \in \{0,1,2\}^d$ with $\max_j p_j = 2$ (which holds when $\boldsymbol{\theta} \in T$ and $\mathbf{m} = (n_1, \ldots, n_d)$), then $V(\boldsymbol{\theta})$ can be expressed as

$$V(\boldsymbol{\theta}) = \sum_{\mathbf{q} \in \{0,1\}^d \setminus \{\mathbf{0}\}} \left[\prod_{j \in S_{\mathbf{q}}} m_j \cdot \Big((D^{(\mathbf{q})}\theta)_{(1, j \in S_{\mathbf{q}}) \times (0, j \notin S_{\mathbf{q}})} - (D^{(\mathbf{q})}\theta)_{(m_j - 1, j \in S_{\mathbf{q}}) \times (0, j \notin S_{\mathbf{q}})} \Big) \right], \tag{D.21}$$

where $S_{\mathbf{q}} := \{j \in [d] : q_j = 1\}$. Here, for each $\mathbf{q}$, $(1, j \in S_{\mathbf{q}}) \times (0, j \notin S_{\mathbf{q}})$ represents the d-dimensional vector whose j^{th} component is 1 if $j \in S_{\mathbf{q}}$, 0 otherwise; $(m_j - 1, j \in S_{\mathbf{q}}) \times (0, j \notin S_{\mathbf{q}})$ is similarly defined. Moreover, for each $V > 0$, define

$$C_{\mathbf{m}}^{d,s}(V) = \Big\{ \boldsymbol{\theta} \in \mathbb{R}^{|I(\mathbf{m})|} : V(\boldsymbol{\theta}) \leq V \text{ and } (D^{(\mathbf{p})}\theta)_{\mathbf{i}} = 0$$
$$\text{for every } \mathbf{p} \in \{0,1\}^d \text{ with } \sum_{j=1}^{d} p_j > s \text{ and } \mathbf{i} \in I(\mathbf{m}) \text{ with } i_j \geq p_j \text{ for } j \in [d] \Big\}.$$

Theorem 29. *There exists some constant C_d depending only on d such that, for every $t > 0$,*

$$\begin{aligned} \mathbb{E}\Big[\sup_{\boldsymbol{\theta} \in C_{\mathbf{m}}^{d,s}(V) : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq t} \langle \boldsymbol{\xi}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \Big] &\leq C_d \sigma t \log(m_1 \cdots m_d) + C_d \sigma t \log\left(1 + \frac{V}{t}\right) \\ &\quad + C_d \sigma t \Big[\log\left(2 + \frac{V (m_1 \cdots m_d)^{1/2}}{t}\right) \Big]^{3(2s-1)/8} \\ &\quad + C_d \sigma V^{1/4} t^{3/4} (m_1 \cdots m_d)^{1/8} \Big[\log\left(2 + \frac{V (m_1 \cdots m_d)^{1/2}}{t}\right) \Big]^{3(2s-1)/8} \end{aligned}$$

when $s \geq 2$, and

$$\mathbb{E}\Big[\sup_{\boldsymbol{\theta} \in C_{\mathbf{m}}^{d,1}(V) : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq t} \langle \boldsymbol{\xi}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \Big]$$

$$\leq C_d \sigma t \log(m_1 \cdots m_d) + C_d \sigma t \log\left(1 + \frac{V}{t}\right) + C_d \sigma V^{1/4} t^{3/4} (m_1 \cdots m_d)^{1/8}$$

when $s = 1$.

We will apply this theorem to the subvectors $\boldsymbol{\theta}^{(\mathbf{r})}$. For each $\mathbf{r} \in \mathcal{R}$, define

$$m_{r_j}^{(j)} := |L_{r_j}^{(j)}| = b_{r_j}^{(j)} - a_{r_j}^{(j)} + 1 \quad \text{for } j \in [d],$$

and let $\mathbf{m}_\mathbf{r} = (m_{r_1}^{(1)}, \dots, m_{r_d}^{(d)})$. We re-index each $\boldsymbol{\theta}^{(\mathbf{r})}$ with $I(\mathbf{m}_\mathbf{r})$ by shifting indices, ensuring that the definition of $V(\cdot)$ is applicable to $\boldsymbol{\theta}^{(\mathbf{r})}$. We will show that, for each $\mathbf{r} \in \mathcal{R}$, there exists some $V_\mathbf{r} > 0$ such that $\boldsymbol{\theta}^{(\mathbf{r})} \in C_{\mathbf{m}_\mathbf{r}}^{d,s}(V_\mathbf{r})$ for every $\boldsymbol{\theta} \in T$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq t$.

Fix $\boldsymbol{\theta} \in T$ and assume that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq t$. Since $\boldsymbol{\theta} \in T$, it follows that $(D^{(\mathbf{p})}\theta)_\mathbf{i} = 0$ for every $\mathbf{p} \in \{0,1\}^d$ with $\sum_{j=1}^d p_j > s$ and $\mathbf{i} = (i_1, \dots, i_d) \in I_0$ with $i_j \geq p_j$ for $j \in [d]$. Thus, for every $\mathbf{p} \in \{0,1\}^d$ with $\sum_{j=1}^d p_j > s$ and $\mathbf{i} = (i_1, \dots, i_d) \in I(\mathbf{m}_\mathbf{r})$ with $i_j \geq p_j$ for $j \in [d]$,

$$(D^{(\mathbf{p})}\theta^{(\mathbf{r})})_\mathbf{i} = (D^{(\mathbf{p})}\theta)_{i_1 + a_{r_1}^{(1)}, \dots, i_d + a_{r_d}^{(d)}} = 0.$$

Next, we bound $V(\boldsymbol{\theta}^{(\mathbf{r})})$ for each $\mathbf{r} \in \mathcal{R}$. To this end, we use the following lemma, which will be proved in Appendix D.13. Choose any function $g^* \in \mathcal{F}_{\mathrm{tc}}^{d,s}$ that agrees with f^* at the design points $\mathbf{x}^{(\mathbf{i})}$, and define

$$M := M(g^*) := \max_{0 < |S| \leq s} \max\left(\left| \lim_{t \to 0+} \begin{bmatrix} 0, t, & j \in S \\ 0, & j \notin S \end{bmatrix} ; \ g^* \right] \right|, \ \lim_{t \to 1-} \left| \begin{bmatrix} t, 1, & j \in S \\ 0, & j \notin S \end{bmatrix} ; \ g^* \right] \right| \right). \tag{D.22}$$

Since $g^* \in \mathcal{F}_{\mathrm{tc}}^{d,s}$, it holds that $M(g^*) < \infty$. Also, note that $\boldsymbol{\theta}^* = (f^*(\mathbf{x}^{(\mathbf{i})}), \mathbf{i} \in I_0) = (g^*(\mathbf{x}^{(\mathbf{i})}), \mathbf{i} \in I_0)$.

Lemma 31. *Assume that* $\boldsymbol{\theta} \in T$ *and* $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq t$. *Then, for every* $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1\}^d \setminus \{\mathbf{0}\}$ *and* $\mathbf{i} = (i_1, \dots, i_d) \in L_\mathbf{r}$ *with* $i_j \geq p_j$ *for* $j \in [d]$, *we have*

$$(D^{(\mathbf{p})}\theta)_\mathbf{i} \geq \frac{C_d}{\prod_{j=1}^d n_j^{p_j}} \cdot \left[-M - t\left(\frac{2^{r_+}}{n}\right)^{1/2} \cdot 2^{\sum_{j=1}^d p_j r_j} \right] \tag{D.23}$$

if $\mathbf{r} = (r_1, \dots, r_d) \in \prod_{j=1}^d \{1, \dots, R_j\} = \mathcal{R}$, *and*

$$(D^{(\mathbf{p})}\theta)_\mathbf{i} \leq \frac{C_d}{\prod_{j=1}^d n_j^{p_j}} \cdot \left[M + t\left(\frac{2^{r_+}}{n}\right)^{1/2} \cdot 2^{\sum_{j=1}^d p_j r_j} \right] \tag{D.24}$$

if $\mathbf{r} = (r_1, \dots, r_d) \in \prod_{j=1}^d \{0, 1, \dots, R_j\}$, *where* $r_+ = \sum_{j=1}^d r_j$.

Recall that $S_{\mathbf{p}} = \{j \in [d] : p_j = 1\}$ for each $\mathbf{p} \in \{0,1\}^d$. By Lemma 31 and (D.21), for each $\mathbf{r} = (r_1, \dots, r_d) \in \mathcal{R}$, we have

$$\begin{aligned}
V(\boldsymbol{\theta}^{(\mathbf{r})}) &= \sum_{\mathbf{p}\in\{0,1\}^d\setminus\{\mathbf{0}\}} \prod_{j\in S_{\mathbf{p}}} m_{r_j}^{(j)} \cdot \Big[(D^{(\mathbf{p})}\theta^{(\mathbf{r})})_{(1,j\in S_{\mathbf{p}})\times(0,j\notin S_{\mathbf{p}})} - (D^{(\mathbf{p})}\theta^{(\mathbf{r})})_{(m_{r_j}^{(j)}-1,j\in S_{\mathbf{p}})\times(0,j\notin S_{\mathbf{p}})}\Big] \\
&= \sum_{\mathbf{p}\in\{0,1\}^d\setminus\{\mathbf{0}\}} \prod_{j\in S_{\mathbf{p}}} m_{r_j}^{(j)} \cdot \Big[(D^{(\mathbf{p})}\theta)_{(a_{r_j}^{(j)}+1,j\in S_{\mathbf{p}})\times(a_{r_j}^{(j)},j\notin S_{\mathbf{p}})} - (D^{(\mathbf{p})}\theta)_{(b_{r_j}^{(j)},j\in S_{\mathbf{p}})\times(a_{r_j}^{(j)},j\notin S_{\mathbf{p}})}\Big] \\
&\le \sum_{\mathbf{p}\in\{0,1\}^d\setminus\{\mathbf{0}\}} \prod_{j=1}^d \Big(\frac{n_j}{2^{r_j}}\Big)^{p_j} \cdot \frac{C_d}{\prod_{j=1}^d n_j^{p_j}} \cdot \Big[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^d p_j r_j}\Big] \\
&\le \sum_{\mathbf{p}\in\{0,1\}^d\setminus\{\mathbf{0}\}} C_d\Big[\frac{M}{2^{\sum_{j=1}^d p_j r_j}} + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2}\Big] \le C_d\Big[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2}\Big] =: V_{\mathbf{r}}.
\end{aligned}$$

Here, for the first inequality, we use

$$m_{r_j}^{(j)} = b_{r_j}^{(j)} - a_{r_j}^{(j)} + 1 \le \frac{n_j}{2^{r_j}} - 1 + 1 = \frac{n_j}{2^{r_j}}$$

for $j \in [d]$. Thus, we have $\boldsymbol{\theta}^{(\mathbf{r})} \in C_{\mathbf{m_r}}^{d,s}(V_{\mathbf{r}})$ for each $\mathbf{r} \in \mathcal{R}$.

From now on, we handle the two cases $s \ge 2$ and $s = 1$ separately. Let us first consider the case $s \ge 2$. By Theorem 29, for each $\mathbf{r} = (r_1, \dots, r_d) \in \mathcal{R}$, we have

$$\begin{aligned}
\mathbb{E}\Big[\sup_{\substack{\boldsymbol{\theta}\in T:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le t \\ \|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2\le t(w_{\mathbf{r}}/|\mathcal{R}|)^{1/2}}} \langle \boldsymbol{\xi}^{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})} - \boldsymbol{\theta}^{*(\mathbf{r})}\rangle\Big] &\le \mathbb{E}\Big[\sup_{\substack{\boldsymbol{\theta}^{(\mathbf{r})}\in C_{\mathbf{m_r}}^{d,s}(V_{\mathbf{r}}) \\ \|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2\le t_{\mathbf{r}}}} \langle \boldsymbol{\xi}^{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})} - \boldsymbol{\theta}^{*(\mathbf{r})}\rangle\Big] \\
&\le C_d\sigma t_{\mathbf{r}} \log|L_{\mathbf{r}}| + C_d\sigma t_{\mathbf{r}} \log\Big(1 + \frac{V_{\mathbf{r}}}{t_{\mathbf{r}}}\Big) + C_d\sigma t_{\mathbf{r}}\Big[\log\Big(2 + \frac{V_{\mathbf{r}}|L_{\mathbf{r}}|^{1/2}}{t_{\mathbf{r}}}\Big)\Big]^{3(2s-1)/8} \\
&\quad + C_d\sigma V_{\mathbf{r}}^{1/4} t_{\mathbf{r}}^{3/4}|L_{\mathbf{r}}|^{1/8}\Big[\log\Big(2 + \frac{V_{\mathbf{r}}|L_{\mathbf{r}}|^{1/2}}{t_{\mathbf{r}}}\Big)\Big]^{3(2s-1)/8}, \qquad (D.25)
\end{aligned}$$

where $t_{\mathbf{r}} = t(w_{\mathbf{r}}/|\mathcal{R}|)^{1/2}$. Recall that

$$\mathbb{E}\Big[\sup_{\boldsymbol{\theta}\in\Lambda_{\mathbf{w}}(t)} \sum_{\mathbf{i}\in J_{\mathbf{0}}} \xi_{\mathbf{i}}(\theta_{\mathbf{i}} - \theta_{\mathbf{i}}^*)\Big] \le \sum_{\mathbf{r}\in\mathcal{R}} \mathbb{E}\Big[\sup_{\substack{\boldsymbol{\theta}\in T:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le t \\ \|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2\le t(w_{\mathbf{r}}/|\mathcal{R}|)^{1/2}}} \langle \boldsymbol{\xi}^{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})} - \boldsymbol{\theta}^{*(\mathbf{r})}\rangle\Big].$$

Hence, we need to bound the summation of each term in (D.25) over $\mathcal{R}$. The summation of the first term over $\mathbf{r} \in \mathcal{R}$ can be bounded as

$$\begin{aligned}
\sum_{\mathbf{r}\in\mathcal{R}} t_{\mathbf{r}} \log|L_{\mathbf{r}}| &\le \sum_{\mathbf{r}\in\mathcal{R}} t\Big(\frac{w_{\mathbf{r}}}{|\mathcal{R}|}\Big)^{1/2} \log\Big(\prod_{j=1}^d \frac{n_j}{2^{r_j}}\Big) \le \frac{t\log n}{|\mathcal{R}|^{1/2}} \cdot \sum_{\mathbf{r}\in\mathcal{R}} w_{\mathbf{r}}^{1/2} \le t\log n \cdot \Big(\sum_{\mathbf{r}\in\mathcal{R}} w_{\mathbf{r}}\Big)^{1/2} \\
&\le t\log n(2|\mathcal{R}|)^{1/2} \le C_d t(\log n)^{(d+2)/2}.
\end{aligned}$$

Here, the first inequality follows from the fact that

$$|L_{\mathbf{r}}| = \prod_{j=1}^{d} |L_{r_j}^{(j)}| = \prod_{j=1}^{d} m_{r_j}^{(j)} \le \prod_{j=1}^{d} \frac{n_j}{2^{r_j}}.$$

Also, the second-to-last inequality is due to the condition $\sum_{\mathbf{r}\in\mathcal{R}} w_{\mathbf{r}} \le 2|\mathcal{R}|$, and the last one follows from

$$|\mathcal{R}| = \prod_{j=1}^{d} R_j \le \prod_{j=1}^{d}(1 + C\log n_j) \le \left[\frac{1}{d}\sum_{j=1}^{d}(1 + C\log n_j)\right]^d = \Big(1 + \frac{C}{d}\cdot\log n\Big)^d \le C_d(\log n)^d.$$

We can bound the summation of the second term over $\mathbf{r} \in \mathcal{R}$ as follows:

$$\begin{aligned}
\sum_{\mathbf{r}\in\mathcal{R}} t_{\mathbf{r}} \log\Big(1 + \frac{V_{\mathbf{r}}}{t_{\mathbf{r}}}\Big) &= \sum_{\mathbf{r}\in\mathcal{R}} t\Big(\frac{w_{\mathbf{r}}}{|\mathcal{R}|}\Big)^{1/2} \cdot \log\Big(1 + C_d \cdot \frac{M + t(2^{r_+}/n)^{1/2}}{t(w_{\mathbf{r}}/|\mathcal{R}|)^{1/2}}\Big) \\
&\le \sum_{\mathbf{r}\in\mathcal{R}} t\Big(\frac{w_{\mathbf{r}}}{|\mathcal{R}|}\Big)^{1/2} \cdot \log\Big(1 + C_d \cdot \frac{(M+t)|\mathcal{R}|^{1/2}}{t}\Big) \\
&\le C_d t(\log n)^{d/2} \cdot \log\Big(1 + C_d \cdot \frac{(M+t)(\log n)^{d/2}}{t}\Big) \\
&\le C_d t(\log n)^{d/2}\Big[C_d + \log\Big(1 + \frac{M}{t}\Big) + \frac{d}{2}\log\log n\Big] \\
&\le C_d t(\log n)^{d/2} + C_d t(\log n)^{d/2}\log\Big(1 + \frac{M}{t}\Big) + C_d t(\log n)^{d/2}\log\log n,
\end{aligned}$$

where the first inequality follows from the facts that $2^{r_+} \le n$ and $w_{\mathbf{r}} \ge 1$. Similarly, the summation of the third term over $\mathbf{r} \in \mathcal{R}$ can be bounded as

$$\begin{aligned}
&\sum_{\mathbf{r}\in\mathcal{R}} t_{\mathbf{r}}\Big[\log\Big(2 + \frac{V_{\mathbf{r}}|L_{\mathbf{r}}|^{1/2}}{t_{\mathbf{r}}}\Big)\Big]^{3(2s-1)/8} \\
&\quad\le \sum_{\mathbf{r}\in\mathcal{R}} t\Big(\frac{w_{\mathbf{r}}}{|\mathcal{R}|}\Big)^{1/2}\Big[\log\Big(2 + C_d \cdot \frac{M + t(2^{r_+}/n)^{1/2}}{t\cdot(w_{\mathbf{r}}/|\mathcal{R}|)^{1/2}} \cdot n^{1/2}\Big)\Big]^{3(2s-1)/8} \\
&\quad\le C_d t(\log n)^{d/2}\Big[\log\Big(2 + C_d \cdot \frac{(M+t)n^{1/2}(\log n)^{d/2}}{t}\Big)\Big]^{3(2s-1)/8} \\
&\quad\le C_d t(\log n)^{d/2}\Big[C_d + \log\Big(1 + \frac{M}{t}\Big) + \frac{1}{2}\log n + \frac{d}{2}\log\log n\Big]^{3(2s-1)/8} \\
&\quad\le C_d t(\log n)^{d/2}\Big[\log\Big(1 + \frac{M}{t}\Big)\Big]^{3(2s-1)/8} + C_d t(\log n)^{(4d+6s-3)/8},
\end{aligned}$$

where the first inequality uses the fact that $|L_{\mathbf{r}}| \le n$. Lastly, the summation of the fourth term over $\mathbf{r} \in \mathcal{R}$ has the following bound:

$$\sum_{\mathbf{r}\in\mathcal{R}} V_{\mathbf{r}}^{1/4} t_{\mathbf{r}}^{3/4} |L_{\mathbf{r}}|^{1/8}\Big[\log\Big(2 + \frac{V_{\mathbf{r}}|L_{\mathbf{r}}|^{1/2}}{t_{\mathbf{r}}}\Big)\Big]^{3(2s-1)/8}$$

$$
\begin{aligned}
&\le \sum_{\mathbf{r}\in\mathcal{R}} C_d\Big[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2}\Big]^{1/4}\Big(t\Big(\frac{w_{\mathbf{r}}}{|\mathcal{R}|}\Big)^{1/2}\Big)^{3/4}\Big(\prod_{j=1}^{d}\frac{n_j}{2^{r_j-1}}\Big)^{1/8}\Big[\log\Big(1+\frac{M}{t}\Big)+\log n\Big]^{3(2s-1)/8} \\
&\le C_d \sum_{\mathbf{r}\in\mathcal{R}}\Big[M^{1/4} + t^{1/4}\Big(\frac{2^{r_+}}{n}\Big)^{1/8}\Big]\cdot t^{3/4}\Big(\frac{w_{\mathbf{r}}}{|\mathcal{R}|}\Big)^{3/8}\Big(\frac{n^{1/8}}{2^{r_+/8}}\Big)\Big[\log\Big(1+\frac{M}{t}\Big)+\log n\Big]^{3(2s-1)/8} \\
&\le C_d\Big[M^{1/4}t^{3/4}n^{1/8}\sum_{\mathbf{r}\in\mathcal{R}}\frac{1}{2^{r_+/8}} + t\sum_{\mathbf{r}\in\mathcal{R}}\Big(\frac{w_{\mathbf{r}}}{|\mathcal{R}|}\Big)^{3/8}\Big]\cdot\Big[\log\Big(1+\frac{M}{t}\Big)+\log n\Big]^{3(2s-1)/8}.
\end{aligned}
$$

Here, the second inequality is based on $(x+y)^{1/4} \le x^{1/4} + y^{1/4}$ for $x, y \ge 0$. Observe that

$$
\sum_{\mathbf{r}\in\mathcal{R}}\frac{1}{2^{r_+/8}} = \prod_{j=1}^{d}\sum_{r_j=1}^{R_j}\frac{1}{2^{r_j/8}} \le C.
$$

Also, by Hölder inequality, we have

$$
\sum_{\mathbf{r}\in\mathcal{R}}\Big(\frac{w_{\mathbf{r}}}{|\mathcal{R}|}\Big)^{3/8} \le \frac{1}{|\mathcal{R}|^{3/8}}\cdot\Big(\sum_{\mathbf{r}\in\mathcal{R}} w_{\mathbf{r}}\Big)^{3/8}\Big(\sum_{\mathbf{r}\in\mathcal{R}} 1\Big)^{5/8} \le C|\mathcal{R}|^{5/8} \le C_d(\log n)^{5d/8}.
$$

It thus follows that

$$
\begin{aligned}
&\sum_{\mathbf{r}\in\mathcal{R}} V_{\mathbf{r}}^{1/4}t_{\mathbf{r}}^{3/4}|L_{\mathbf{r}}|^{1/8}\Big[\log\Big(2+\frac{V_{\mathbf{r}}|L_{\mathbf{r}}|^{1/2}}{t_{\mathbf{r}}}\Big)\Big]^{3(2s-1)/8} \\
&\qquad \le C_d M^{1/4}t^{3/4}n^{1/8}\Big[\log\Big(1+\frac{M}{t}\Big)\Big]^{3(2s-1)/8} + C_d M^{1/4}t^{3/4}n^{1/8}(\log n)^{3(2s-1)/8} \\
&\qquad\quad + C_d t(\log n)^{5d/8}\Big[\log\Big(1+\frac{M}{t}\Big)\Big]^{3(2s-1)/8} + C_d t(\log n)^{(5d+6s-3)/8}.
\end{aligned}
$$

Combining these results, we can derive that

$$
\begin{aligned}
\mathbb{E}\Big[\sup_{\boldsymbol{\theta}\in\Lambda_{\mathbf{w}}(t)}\sum_{\mathbf{i}\in J_{\mathbf{0}}}\xi_{\mathbf{i}}(\theta_{\mathbf{i}}-\theta^*_{\mathbf{i}})\Big] &\le C_d\sigma t(\log n)^{\max((d+2)/2,(5d+6s-3)/8)} + C_d\sigma t(\log n)^{d/2}\log\Big(1+\frac{M}{t}\Big) \\
&\quad + C_d\sigma t(\log n)^{5d/8}\Big[\log\Big(1+\frac{M}{t}\Big)\Big]^{3(2s-1)/8} + C_d\sigma M^{1/4}t^{3/4}n^{1/8}\Big[\log\Big(1+\frac{M}{t}\Big)\Big]^{3(2s-1)/8} \\
&\quad + C_d\sigma M^{1/4}t^{3/4}n^{1/8}(\log n)^{3(2s-1)/8}.
\end{aligned}
$$

Using (D.20), we can show that $H_{\mathbf{0}}(t)$ has the same upper bound (with a slightly larger C_d). Similarly, by analogous arguments, we can derive the same upper bound for $H_{\boldsymbol{\delta}}(t)$ for $\boldsymbol{\delta} \in \{0,1\}^d \setminus \{\mathbf{0}\}$. Therefore, we can see that if we define

$$
t_2 = \max\Big(\sigma, C_d\sigma(\log n)^{\max((d+2)/2,(5d+6s-3)/8)}, C_d\sigma(\log n)^{d/2}\cdot\log\Big(1+\frac{M}{\sigma}\Big),
$$

$$C_d\sigma(\log n)^{5d/8}\Big[\log\Big(1+\frac{M}{\sigma}\Big)\Big]^{3(2s-1)/8}, C_d\sigma^{4/5}M^{1/5}n^{1/10}\Big[\log\Big(1+\frac{M}{\sigma}\Big)\Big]^{3(2s-1)/10},$$
$$C_d\sigma^{4/5}M^{1/5}n^{1/10}(\log n)^{3(2s-1)/10}\Big),$$

then $H(t_2) \le t_2^2/2$, which implies $h(t_2) \le 0$. As a result, by Theorem 28 and Proposition 16,

$$\begin{aligned}
R_F(\hat{f}_n^{d,s}, f^*) &= \frac{1}{n}\cdot\mathbb{E}\big[\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2^2\big] \le \frac{C}{n}\cdot(\sigma^2 + t_2^2)\\
&\le C_d\Big(\frac{\sigma^2 M^{1/2}}{n}\Big)^{4/5}(\log n)^{3(2s-1)/5} + C_d\Big(\frac{\sigma^2 M^{1/2}}{n}\Big)^{4/5}\Big[\log\Big(1+\frac{M}{\sigma}\Big)\Big]^{3(2s-1)/5}\\
&\quad + C_d\cdot\frac{\sigma^2}{n}(\log n)^{5d/4}\Big[\log\Big(1+\frac{M}{\sigma}\Big)\Big]^{3(2s-1)/4} + C_d\cdot\frac{\sigma^2}{n}(\log n)^{d}\Big[\log\Big(1+\frac{M}{\sigma}\Big)\Big]^{2}\\
&\quad + C_d\cdot\frac{\sigma^2}{n}(\log n)^{\max(d+2,(5d+6s-3)/4)}.
\end{aligned}$$

This can be simplified as

$$\begin{aligned}
R_F(\hat{f}_n^{d,s}, f^*) &\le C_d\Big(\frac{\sigma^2 M^{1/2}}{n}\Big)^{4/5}\Big[\log\Big(\Big(1+\frac{M}{\sigma}\Big)n\Big)\Big]^{3(2s-1)/5}\\
&\quad + C_d\cdot\frac{\sigma^2}{n}(\log n)^{5d/4}\Big[\log\Big(\Big(1+\frac{M}{\sigma}\Big)n\Big)\Big]^{3(2s-1)/4},
\end{aligned}\tag{D.26}$$

as desired, since $s \ge 2$.

We now consider the case $s = 1$. In this case, again, by Theorem 29, we have

$$\begin{aligned}
\mathbb{E}\Bigg[\sup_{\substack{\boldsymbol{\theta}\in T:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le t\\ \|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2\le t(w_{\mathbf{r}}/|\mathcal{R}|)^{1/2}}}\langle\boldsymbol{\xi}^{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})} - \boldsymbol{\theta}^{*(\mathbf{r})}\rangle\Bigg] &\le \mathbb{E}\Bigg[\sup_{\substack{\boldsymbol{\theta}^{(\mathbf{r})}\in C^{d,1}_{\mathbf{m}_{\mathbf{r}}}(V_{\mathbf{r}})\\ \|\boldsymbol{\theta}^{(\mathbf{r})}-\boldsymbol{\theta}^{*(\mathbf{r})}\|_2\le t_{\mathbf{r}}}}\langle\boldsymbol{\xi}^{(\mathbf{r})}, \boldsymbol{\theta}^{(\mathbf{r})} - \boldsymbol{\theta}^{*(\mathbf{r})}\rangle\Bigg]\\
&\le C_d\sigma t_{\mathbf{r}}\log|L_{\mathbf{r}}| + C_d\sigma t_{\mathbf{r}}\log\Big(1+\frac{V_{\mathbf{r}}}{t_{\mathbf{r}}}\Big) + C_d\sigma V_{\mathbf{r}}^{1/4}t_{\mathbf{r}}^{3/4}|L_{\mathbf{r}}|^{1/8}.
\end{aligned}$$

Through similar computations as above, we can show that

$$\begin{aligned}
H_{\mathbf{0}}(t) &\le C_d\sigma t(\log n)^{(d+2)/2} + C_d\sigma t(\log n)^{d/2}\log\Big(1+\frac{M}{t}\Big)\\
&\quad + C_d\sigma t(\log n)^{5d/8} + C_d\sigma M^{1/4}t^{3/4}n^{1/8}.
\end{aligned}$$

Also, by analogous arguments, the same upper bound holds for $H_{\boldsymbol{\delta}}(t)$ for $\boldsymbol{\delta} \in \{0,1\}^d \setminus \{\mathbf{0}\}$. Thus, in this case, the inequality $h(t_2) \le 0$ holds for

$$t_2 = \max\Big(\sigma, C_d\sigma(\log n)^{(d+2)/2}, C_d\sigma(\log n)^{d/2}\cdot\log\Big(1+\frac{M}{\sigma}\Big), C_d\sigma(\log n)^{5d/8}, C_d\sigma^{4/5}M^{1/5}n^{1/10}\Big).$$

Consequently, using Theorem 28 and Proposition 16, we can derive that

$$R_F(\hat{f}_n^{d,1}, f^*) = \frac{1}{n}\cdot\mathbb{E}\big[\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2^2\big] \le \frac{C}{n}\cdot(\sigma^2 + t_2^2) \tag{D.27}$$

$$\leq C_d\Big(\frac{\sigma^2 M^{1/2}}{n}\Big)^{4/5} + C_d \cdot \frac{\sigma^2}{n}(\log n)^d\Big[\log\Big(\Big(1+\frac{M}{\sigma}\Big)n\Big)\Big]^2 + C_d \cdot \frac{\sigma^2}{n}(\log n)^{5d/4}.$$

We now show that $M = M(g^*)$ in the bounds (D.26) and (D.27) can be replaced with $V_{\text{design}}(f^*)$. Fix $c_S \in \mathbb{R}$ for each subset $S \subseteq [d]$ with $0 < |S| \leq s$, and let $A : [0,1]^d \to \mathbb{R}$ denote the multi-affine function defined as

$$A(x_1, \dots, x_d) = \sum_{0<|S|\leq s} c_S \prod_{j\in S} x_j \quad \text{for } (x_1, \dots, x_d) \in [0,1]^d.$$

Next, define $\bar{g} = g^* - A \in \mathcal{F}_{\text{tc}}^{d,s}$, and let $\bar{y}_{\mathbf{i}} = y_{\mathbf{i}} - A(\mathbf{x}^{(\mathbf{i})})$ for $\mathbf{i} \in I_0$. It then follows that

$$\bar{y}_{\mathbf{i}} = y_{\mathbf{i}} - A(\mathbf{x}^{(\mathbf{i})}) = f^*(\mathbf{x}^{(\mathbf{i})}) - A(\mathbf{x}^{(\mathbf{i})}) + \xi_{\mathbf{i}} = g^*(\mathbf{x}^{(\mathbf{i})}) - A(\mathbf{x}^{(\mathbf{i})}) + \xi_{\mathbf{i}} = \bar{g}(\mathbf{x}^{(\mathbf{i})}) + \xi_{\mathbf{i}} \tag{D.28}$$

for $\mathbf{i} \in I_0$. We can thus view $(\mathbf{x}^{(\mathbf{i})}, \bar{y}_{\mathbf{i}})$ for $\mathbf{i} \in I_0$ as observations from the model (D.28), where the true regression function is $\bar{g}$. Moreover, define $\bar{f}_n^{d,s} = \hat{f}_n^{d,s} - A$. Since $\mathcal{F}_{\text{tc}}^{d,s}$ is invariant under subtraction by A, i.e.,

$$\mathcal{F}_{\text{tc}}^{d,s} = \big\{f - A : f \in \mathcal{F}_{\text{tc}}^{d,s}\big\},$$

it is clear that $\bar{f}_n^{d,s}$ is a least squares estimator over $\mathcal{F}_{\text{tc}}^{d,s}$, based on observations $(\mathbf{x}^{(\mathbf{i})}, \bar{y}_{\mathbf{i}})$ for $\mathbf{i} \in I_0$. That is,

$$\bar{f}_n^{d,s} \in \operatorname*{argmin}_{f\in\mathcal{F}_{\text{tc}}^{d,s}} \sum_{\mathbf{i}\in I_0} \big(\bar{y}_{\mathbf{i}} - f(\mathbf{x}^{(\mathbf{i})})\big)^2.$$

Hence, the bounds (D.26) and (D.27) hold for $R_F(\bar{f}_n^{d,s}, \bar{g})$ with $M = M(\bar{g})$. Since

$$R_F(\hat{f}_n^{d,s}, f^*) = R_F(\hat{f}_n^{d,s}, g^*) = R_F(\bar{f}_n^{d,s}, \bar{g}),$$

this implies that the bounds (D.26) and (D.27) also hold for $R_F(\hat{f}_n^{d,s}, f^*)$ with $M = M(\bar{g})$.

Observe that, for each subset $S \subseteq [d]$ with $0 < |S| \leq s$,

$$\begin{bmatrix} 0,t, & j\in S \\ 0, & j\notin S \end{bmatrix}; \; \bar{g} \Bigg] = \begin{bmatrix} 0,t, & j\in S \\ 0, & j\notin S \end{bmatrix}; \; g^* \Bigg] - c_S \quad \text{for every } t > 0$$

and

$$\begin{bmatrix} t,1, & j\in S \\ 0, & j\notin S \end{bmatrix}; \; \bar{g} \Bigg] = \begin{bmatrix} t,1, & j\in S \\ 0, & j\notin S \end{bmatrix}; \; g^* \Bigg] - c_S \quad \text{for every } t < 1.$$

Hence, if we choose c_S as

$$c_S = \lim_{t\to 1-} \left[\begin{matrix} t,1, & j\in S \\ 0, & j\notin S \end{matrix} \;;\; g^* \right]$$

for each subset S, then we have

$$M(\bar{g}) = \max_{0<|S|\leq s}\left(\lim_{t\to 0+}\left[\begin{matrix} 0,t, & j\in S \\ 0, & j\notin S \end{matrix} \;;\; g^*\right] - \lim_{t\to 1-}\left[\begin{matrix} t,1, & j\in S \\ 0, & j\notin S \end{matrix} \;;\; g^*\right]\right)$$

(recall the definition of $M(\cdot)$ from (D.22)).

Since $g^* \in \mathcal{F}_{\mathrm{tc}}^{d,s}$, there exist real numbers c_0 and c_S (for $0 < |S| \leq s$) and finite measures ν_S (for $0 < |S| \leq s$) on $[0,1)^{|S|} \setminus \{\mathbf{0}\}$ such that

$$g^*(x_1, \dots, x_d) = c_0 + \sum_{0<|S|\leq s} c_S \prod_{j\in S} x_j - \sum_{0<|S|\leq s} \int_{[0,1)^{|S|}\setminus\{\mathbf{0}\}} \prod_{j\in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S)$$

for $(x_1, \dots, x_d) \in [0,1]^d$. Observe that, for each S, we have

$$\begin{aligned}
\int_{[0,1)^{|S|}\setminus\{\mathbf{0}\}} \prod_{j\in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S) &= \int_{\prod_{j\in S}[0,x_j)\setminus\{\mathbf{0}\}} \int_{\prod_{j\in S}[t_j,x_j)} 1 \, d(s_j, j\in S) \, d\nu_S(t_j, j \in S) \\
&= \int_{\prod_{j\in S}[0,x_j)} \int_{\prod_{j\in S}[0,s_j]\setminus\{\mathbf{0}\}} 1 \, d\nu_S(t_j, j \in S) \, d(s_j, j\in S) \\
&= \int_{\prod_{j\in S}[0,x_j)} \nu_S\Big(\prod_{j\in S}[0,s_j] \setminus \{\mathbf{0}\}\Big) \, d(s_j, j\in S).
\end{aligned}$$

Thus, it can be readily verified that, for each S,

$$\begin{bmatrix} 0,t, & j\in S \\ 0, & j\notin S \end{bmatrix} ; \; g^* \Bigg] = c_S - \frac{1}{t^{|S|}} \int_{\prod_{j\in S}[0,t)} \nu_S\Big(\prod_{j\in S}[0,s_j] \setminus \{\mathbf{0}\}\Big) \, d(s_j, j\in S)$$

for every $t > 0$, and

$$\begin{bmatrix} t,1, & j\in S \\ 0, & j\notin S \end{bmatrix} ; \; g^* \Bigg] = c_S - \frac{1}{(1-t)^{|S|}} \int_{\prod_{j\in S}[t,1)} \nu_S\Big(\prod_{j\in S}[0,s_j] \setminus \{\mathbf{0}\}\Big) \, d(s_j, j\in S)$$

for every $t < 1$. Since ν_S is nonnegative, this implies that

$$\lim_{t\to 0+} \left[\begin{matrix} 0,t, & j\in S \\ 0, & j\notin S \end{matrix} \; ; \; g^* \right] \leq c_S$$

and

$$\lim_{t\to 1-} \left[\begin{matrix} t,1, & j\in S \\ 0, & j\notin S \end{matrix} \; ; \; g^* \right] \geq c_S - \nu_S\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big).$$

As a result, we have

$$M(\bar{g}) \leq \max_{0<|S|\leq s} \nu_S\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big) \leq \sum_{0<|S|\leq s} \nu_S\big([0,1)^{|S|} \setminus \{\mathbf{0}\}\big) = V_{\mathrm{mars}}(g^*).$$

Recall that the bounds (D.26) and (D.27) hold for $R_F(\hat{f}_n^{d,s}, f^*)$ with $M = M(\bar{g})$. Also, recall that g^* is any function in $\mathcal{F}_{\mathrm{tc}}^{d,s}$ that agrees with f^* at the design points $\mathbf{x}^{(\mathbf{i})}$. Hence, the bounds (D.26) and (D.27) still remain valid when M is replaced with $V_{\mathrm{design}}(f^*)$. $\square$

D.6 Proof of Theorem 11

The following theorem, from [7], is a key tool for our proof of Theorem 11. Recall the model (D.13) and the definition (D.14) of the least squares estimator $\hat{\boldsymbol{\theta}}_K$ over a closed convex subset K of $\mathbb{R}^n$. Theorem 30 provides an alternative bound for the expected squared error $\mathbb{E}[\|\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}^*\|_2^2]$ of $\hat{\boldsymbol{\theta}}_K$. The bound in Theorem 30 is expressed in terms of the *statistical dimension* of the *tangent cone* of K, which are defined as follows.

For a closed convex subset K of $\mathbb{R}^n$, the tangent cone of K at $\boldsymbol{\theta} \in K$ is defined as

$$\mathcal{T}_K(\boldsymbol{\theta}) = \overline{\{t(\boldsymbol{\eta} - \boldsymbol{\theta}) : t \geq 0 \text{ and } \boldsymbol{\eta} \in K\}},$$

where $\overline{\{\cdot\}}$ denotes the closure. It is clear that $\mathcal{T}_K(\boldsymbol{\theta})$ is closed and convex. Also, it can be readily checked that $\mathcal{T}_K(\boldsymbol{\theta})$ is a cone, meaning that if $\boldsymbol{\zeta} \in \mathcal{T}_K(\boldsymbol{\theta})$ and $t \geq 0$, then $t\boldsymbol{\zeta} \in \mathcal{T}_K(\boldsymbol{\theta})$. For a closed convex cone $\mathcal{T} \subseteq \mathbb{R}^n$, the statistical dimension of $\mathcal{T}$ is defined by

$$\delta(\mathcal{T}) = \mathbb{E}\|\Pi_{\mathcal{T}}(\mathbf{Z})\|_2^2,$$

where $\mathbf{Z} = (Z_1, \ldots, Z_n)$ is a Gaussian random vector with $Z_i \overset{\text{i.i.d.}}{\sim} N(0, 1^2)$, $\Pi_{\mathcal{T}}(\mathbf{Z})$ denotes the projection of $\mathbf{Z}$ onto $\mathcal{T}$, and $\|\cdot\|_2$ is the Euclidean norm. It is well-known that the statistical dimension $\delta(\cdot)$ is monotone: for closed convex cones $\mathcal{T} \subseteq \mathcal{T}'$, we have $\delta(\mathcal{T}) \leq \delta(\mathcal{T}')$ (see, e.g., [2, Proposition 3.1]).

Theorem 30 (Corollary 2.2 of [7])**.** *Under the model* (D.13)*,* $\hat{\boldsymbol{\theta}}_K$ *(defined in* (D.14)*) satisfies*

$$\mathbb{E}\big[\|\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}^*\|_2^2\big] \leq \inf_{\boldsymbol{\theta} \in K} \big\{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 + \sigma^2 \cdot \delta(\mathcal{T}_K(\boldsymbol{\theta}))\big\}.$$

We also use the following notation, which extends T defined in (D.16), in our proof of Theorem 11. For positive integers $m_1, \ldots, m_d$, we define

$$T(m_1, \ldots, m_d) = \left\{\Big(f\Big(\frac{i_1}{m_1}, \ldots, \frac{i_d}{m_d}\Big), (i_1, \ldots, i_d) \in I(m_1, \ldots, m_d)\Big) : f \in \mathcal{F}_{\text{tc}}^{d,s}\right\},$$

where $I(m_1, \ldots, m_d)$ is the index set defined as in (D.18). Note that $T = T(n_1, \ldots, n_d)$.

Furthermore, as in the proof of Theorem 9, we re-index $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$ as $(\mathbf{x}^{(\mathbf{i})}, y_{\mathbf{i}})$ for $\mathbf{i} \in I_0$ in our proof of Theorem 11. For each real-valued function f on $[0,1]^d$, we also denote by $\boldsymbol{\theta}_f$ the vector of evaluations of f at the design points:

$$\boldsymbol{\theta}_f = (f(\mathbf{x}^{(\mathbf{i})}), \mathbf{i} \in I_0).$$

Proof of Theorem 11. Recall that $\boldsymbol{\theta}^* = (f^*(\mathbf{x}^{(\mathbf{i})}), \mathbf{i} \in I_0) = \boldsymbol{\theta}_{f^*}$, and $\hat{\boldsymbol{\theta}}_T$ is defined as the least squares estimator over T:

$$\hat{\boldsymbol{\theta}}_T = \operatorname*{argmin}_{\boldsymbol{\theta} \in T} \|\mathbf{y} - \boldsymbol{\theta}\|_2,$$

where $\mathbf{y} = (y_{\mathbf{i}}, \mathbf{i} \in I_0)$. Also, recall that we have

$$R_F(\hat{f}_n^{d,s}, f^*) = \frac{1}{n} \cdot \mathbb{E}\big[\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2^2\big].$$

Since $T = \{\boldsymbol{\theta}_f : f \in \mathcal{F}_{\mathrm{tc}}^{d,s}\}$, applying Theorem 30 with $K = T$, we obtain

$$\begin{aligned} R_F(\hat{f}_n^{d,s}, f^*) = \frac{1}{n} \cdot \mathbb{E}\big[\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\|_2^2\big] &\le \inf_{\boldsymbol{\theta} \in T} \Big\{ \frac{1}{n} \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 + \frac{\sigma^2}{n} \cdot \delta(\mathcal{T}_T(\boldsymbol{\theta})) \Big\} \\ &= \inf_{f \in \mathcal{F}_{\mathrm{tc}}^{d,s}} \Big\{ \|f - f^*\|_n^2 + \frac{\sigma^2}{n} \cdot \delta(\mathcal{T}_T(\boldsymbol{\theta}_f)) \Big\} \\ &\le \inf_{f \in \mathcal{F}_{\mathrm{tc}}^{d,s} \cap \mathcal{R}^{d,s}} \Big\{ \|f - f^*\|_n^2 + \frac{\sigma^2}{n} \cdot \delta(\mathcal{T}_T(\boldsymbol{\theta}_f)) \Big\}. \end{aligned}$$

Hence, to prove Theorem 11, it suffices to show that

$$\delta(\mathcal{T}_T(\boldsymbol{\theta}_f)) \le \begin{cases} C_d \cdot m(f) \cdot (\log n)^{(5d+6s-3)/4} & \text{if } s \ge 2, \\ C_d \cdot m(f) \cdot (\log n)^{\max(d+2, 5d/4)} & \text{if } s = 1 \end{cases}$$

for every $f \in \mathcal{F}_{\mathrm{tc}}^{d,s} \cap \mathcal{R}^{d,s}$.

Fix $f \in \mathcal{F}_{\mathrm{tc}}^{d,s} \cap \mathcal{R}^{d,s}$, and let

$$P = \Big\{ \prod_{j=1}^{d} \big[v_{l_j - 1}^{(j)}, v_{l_j}^{(j)}\big] : l_j \in [m_j] \Big\}$$

be an axis-aligned split of $[0,1]^d$ with $|P| = m(f)$ such that $0 = v_0^{(j)} < \cdots < v_{m_j}^{(j)} = 1$ for $j \in [d]$, and f is a multi-affine function of interaction order s on each rectangle of P. For each $\boldsymbol{\theta} \in \mathbb{R}^{|I_0|}$ and $\mathbf{l} = (l_1, \ldots, l_d) \in \prod_{j=1}^{d} [m_j]$, let $\boldsymbol{\theta}^{(\mathbf{l})}$ be the subvector of $\boldsymbol{\theta}$ defined as

$$\boldsymbol{\theta}^{(\mathbf{l})} = \Big(\theta_{i_1, \ldots, i_d}, \Big(\frac{i_1}{n_1}, \ldots, \frac{i_d}{n_d}\Big) \in \prod_{j=1}^{d} \big[v_{l_j - 1}^{(j)}, v_{l_j}^{(j)}\big) \Big).$$

Also, for each $\mathbf{l}$, by shifting indices, we re-index $\boldsymbol{\theta}^{(\mathbf{l})}$ with $I(n_1^{(\mathbf{l})}, \ldots, n_d^{(\mathbf{l})})$, where $n_j^{(\mathbf{l})}$ is the number of i_j's for which $i_j / n_j \in [v_{l_j - 1}^{(j)}, v_{l_j}^{(j)})$.

Now, we claim that

$$\mathcal{T}_T(\boldsymbol{\theta}_f) \subseteq \Big\{ \boldsymbol{\theta} \in \mathbb{R}^{|I_0|} : \boldsymbol{\theta}^{(\mathbf{l})} \in T\big(n_1^{(\mathbf{l})}, \ldots, n_d^{(\mathbf{l})}\big) \text{ for every } \mathbf{l} = (l_1, \ldots, l_d) \in \prod_{j=1}^{d} [m_j] \Big\} =: \mathcal{T}_P.$$

It is clear that $\mathcal{T}_P$ is closed, and if $\boldsymbol{\theta} \in \mathcal{T}_P$ and $t \ge 0$, then $t\boldsymbol{\theta} \in \mathcal{T}_P$. For the claim, it is thus enough to show that, for every $\boldsymbol{\theta} \in T$, we have $\boldsymbol{\theta} - \boldsymbol{\theta}_f \in \mathcal{T}_P$. Fix $\boldsymbol{\theta} \in T$, and choose any function $g \in \mathcal{F}_{\text{tc}}^{d,s}$ with $\boldsymbol{\theta}_g = \boldsymbol{\theta}$. Also, fix $\mathbf{l} = (l_1, \ldots, l_d) \in \prod_{j=1}^d [m_j]$, and, for each $j \in [d]$, let $k_j^{(\mathbf{l})}$ be the smallest integer for which $k_j^{(\mathbf{l})}/n_j \in [v_{l_j-1}^{(j)}, v_{l_j}^{(j)})$. For these fixed $\boldsymbol{\theta}$ and $\mathbf{l}$, we show that $(\boldsymbol{\theta} - \boldsymbol{\theta}_f)^{(\mathbf{l})} = \boldsymbol{\theta}_{g-f}^{(\mathbf{l})} \in T(n_1^{(\mathbf{l})}, \ldots, n_d^{(\mathbf{l})})$, from which the claim follows directly.

Suppose f is of the form

$$f(x_1, \ldots, x_d) = c_0^{(\mathbf{l})} + \sum_{0 < |S| \leq s} c_S^{(\mathbf{l})} \prod_{j \in S} x_j \tag{D.29}$$

on $\prod_{j=1}^d [v_{l_j-1}^{(j)}, v_{l_j}^{(j)}]$, for some constants $c_0^{(\mathbf{l})}$ and $c_S^{(\mathbf{l})}$, $0 < |S| \leq s$. First, let $f_0 : [0,1]^d \to \mathbb{R}$ be the function that is of the form (D.29) over the entire domain $[0,1]^d$. Note that while f may not be a multi-affine function on $\prod_{j=1}^d [k_j^{(\mathbf{l})}/n_j, (k_j^{(\mathbf{l})} + n_j^{(\mathbf{l})})/n_j]$, the function f_0 is multi-affine on $\prod_{j=1}^d [k_j^{(\mathbf{l})}/n_j, (k_j^{(\mathbf{l})} + n_j^{(\mathbf{l})})/n_j]$ by definition. Next, define h as

$$h(x_1, \ldots, x_d) = (g - f_0)\Big(\frac{k_1^{(\mathbf{l})} + n_1^{(\mathbf{l})} \cdot x_1}{n_1}, \ldots, \frac{k_d^{(\mathbf{l})} + n_d^{(\mathbf{l})} \cdot x_d}{n_d}\Big)$$

for $(x_1, \ldots, x_d) \in [0,1]^d$. Observe that

$$\Big(\frac{k_1^{(\mathbf{l})} + n_1^{(\mathbf{l})} \cdot x_1}{n_1}, \ldots, \frac{k_d^{(\mathbf{l})} + n_d^{(\mathbf{l})} \cdot x_d}{n_d}\Big) \in \prod_{j=1}^d \big[v_{l_j-1}^{(j)}, v_{l_j}^{(j)}\big)$$

provided that $x_j \leq (n_j^{(\mathbf{l})} - 1)/n_j^{(\mathbf{l})}$ for $j \in [d]$. It thus follows that

$$\boldsymbol{\theta}_{g-f}^{(\mathbf{l})} = \boldsymbol{\theta}_{g-f_0}^{(\mathbf{l})} = \bigg(h\Big(\frac{i_1}{n_1^{(\mathbf{l})}}, \ldots, \frac{i_d}{n_d^{(\mathbf{l})}}\Big), (i_1, \ldots, i_d) \in I\big(n_1^{(\mathbf{l})}, \ldots, n_d^{(\mathbf{l})}\big)\bigg).$$

We now verify that $h \in \mathcal{F}_{\text{tcp}}^d$ and satisfies the interaction restriction condition (4.16) for all subsets $S \subseteq [d]$ with $|S| > s$. By the definition of $T(n_1^{(\mathbf{l})}, \ldots, n_d^{(\mathbf{l})})$ and (the proof of) Lemma 3, this implies that $(\boldsymbol{\theta} - \boldsymbol{\theta}_f)^{(\mathbf{l})} = \boldsymbol{\theta}_{g-f}^{(\mathbf{l})} \in T(n_1^{(\mathbf{l})}, \ldots, n_d^{(\mathbf{l})})$, which proves the claim. To see this, first, note that the divided differences of f_0 of order $\mathbf{p} = (p_1, \ldots, p_d)$ with $\max_j p_j = 2$ are all zero because f_0 is a multi-affine function. Combining this with the fact that $g \in \mathcal{F}_{\text{tc}}^{d,s} \subseteq \mathcal{F}_{\text{tcp}}^d$, it follows that $h \in \mathcal{F}_{\text{tcp}}^d$. Also, it is straightforward to see that the interaction restriction condition (4.16) for h can be derived from those for g and f_0, which are from the fact that $g, f_0 \in \mathcal{F}_{\text{tc}}^{d,s}$.

Due to the monotonicity of the statistical dimension $\delta(\cdot)$, we have

$$\delta(\mathcal{T}_T(\boldsymbol{\theta}_f)) \leq \delta(\mathcal{T}_P) = \mathbb{E}\|\Pi_{\mathcal{T}_P}(\mathbf{Z})\|_2^2 = \sum_{\mathbf{l} \in \prod_{j=1}^d [m_j]} \mathbb{E}\|\Pi_{T(n_1^{(\mathbf{l})}, \ldots, n_d^{(\mathbf{l})})}(\mathbf{Z}^{(\mathbf{l})})\|_2^2.$$

Note that $\mathbb{E}\|\Pi_{T(n_1^{(\mathbf{l})}, \ldots, n_d^{(\mathbf{l})})}(\mathbf{Z}^{(\mathbf{l})})\|_2^2/(n_1^{(\mathbf{l})} \cdots n_d^{(\mathbf{l})})$ is equal to the risk $R_F(\hat{f}_n^{d,s}, f^*)$ of the estimator $\hat{f}_n^{d,s}$, when $f^* = 0$, $\sigma = 1$, and the design points are $(i_1/n_1^{(\mathbf{l})}, \ldots, i_d/n_d^{(\mathbf{l})})$ for

$(i_1, \ldots, i_d) \in I(n_1^{(\mathbf{l})}, \ldots, n_d^{(\mathbf{l})})$. Hence, we can reuse the results established in Theorem 9 to derive that

$$\delta(\mathcal{T}_T(\boldsymbol{\theta}_f)) \leq \sum_{\mathbf{l} \in \prod_{j=1}^d [m_j]} C_d \Big(\log \big(n_1^{(\mathbf{l})} \cdots n_d^{(\mathbf{l})} \big) \Big)^{(5d+6s-3)/4} \leq C_d \cdot m(f) \cdot (\log n)^{(5d+6s-3)/4}$$

when $s \geq 2$,

$$\delta(\mathcal{T}_T(\boldsymbol{\theta}_f)) \leq \sum_{\mathbf{l} \in \prod_{j=1}^d [m_j]} C_d \Big(\log \big(n_1^{(\mathbf{l})} \cdots n_d^{(\mathbf{l})} \big) \Big)^{\max(d+2, 5d/4)} \leq C_d \cdot m(f) \cdot (\log n)^{\max(d+2, 5d/4)}$$

when $s = 1$. This concludes the proof. □

D.7 Proof of Theorem 12

Our proof of Theorem 12 almost follows that of Theorem 5 in Appendix C. As in that proof, Theorem 24, which is a version of [69, Proposition 2], plays a central role in our proof of Theorem 12.

Proof of Theorem 12. For this proof, we use results established for the function class $\mathcal{F}^{d,s}_{\infty-\text{mars}}$ in Chapter 3 and Appendix C. Recall that $\mathcal{F}^{d,s}_{\infty-\text{mars}}$ denotes the collection of all functions of the form (4.1), where for each subset $S \subseteq [d]$ with $0 < |S| \leq s$, ν_S is a finite signed (Borel) measure on $[0,1)^{|S|} \setminus \{\mathbf{0}\}$.

First, using the same argument as in the proof of Lemma 15 in Appendix C, we can prove the following:

Lemma 32. *The constrained variant $\hat{f}^{d,s}_{n,V}$ satisfies*

$$\|\hat{f}^{d,s}_{n,V} - f^*\|_\infty = O_p(1).$$

Fix $\epsilon > 0$. This lemma guarantees that there exists $M > 0$ such that

$$\mathbb{P}\big(\|\hat{f}^{d,s}_{n,V} - f^*\|_\infty \leq M\big) \geq 1 - \epsilon \tag{D.30}$$

for sufficiently large n. Define

$$\mathcal{F}_{\text{tc}}(V, M) = \big\{ f \in \mathcal{F}^{d,s}_{\text{tc}} : V_{\text{mars}}(f) \leq V \text{ and } \|f - f^*\|_\infty \leq M \big\},$$

and

$$\mathcal{F}_{\text{mars}}(V, M) = \big\{ f \in \mathcal{F}^{d,s}_{\infty-\text{mars}} : V_{\text{mars}}(f) \leq V \text{ and } \|f - f^*\|_\infty \leq M \big\}.$$

It is clear that $\mathcal{F}_{\text{tc}}(V, M)$ is a collection of continuous functions on $[0,1]^d$ and contains f^*. Also, from (D.30) and the definition of $\hat{f}^{d,s}_{n,V}$, it follows that, for sufficiently large n,

$$\hat{f}^{d,s}_{n,V} \in \operatorname*{argmin}_f \left\{ \sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)}) \big)^2 : f \in \mathcal{F}_{\text{tc}}(V, M) \right\}$$

with probability at least $1-\epsilon$.

Since $\mathcal{F}_{\mathrm{tc}}^{d,s} \subseteq \mathcal{F}_{\infty-\mathrm{mars}}^{d,s}$, we have $\mathcal{F}_{\mathrm{tc}}(V,M) \subseteq \mathcal{F}_{\mathrm{mars}}(V,M)$. It follows that, for every $t>0$,

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_{\mathrm{tc}}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le t}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] \le \mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_{\mathrm{mars}}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le t}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big]$$

and

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_{\mathrm{tc}}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le t}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] \le \mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_{\mathrm{mars}}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le t}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big],$$

where ϵ_i are i.i.d. Rademacher variables independent of $\mathbf{x}^{(i)}$. Thus, by repeating the arguments in the proof of Theorem 5, we can obtain

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_{\mathrm{tc}}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le rt_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] \le r\sqrt{n}t_n^2$$

and

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_{\mathrm{tc}}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le rt_n}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] \le r\sqrt{n}t_n^2$$

for every $r\ge 1$, provided that

$$\begin{aligned} t_n = (1+4\|\xi_1\|_{5,1})\Big[&C_{B,d}(1+M^{1/2})^{4/5}\big[(1+M^{1/2})^{1/5}+(M+V)^{1/5}\big] \\ &+ C_{B,d}M^{4/9}V^{1/9}\Big[1+\Big(\frac{M+V}{1+M^{1/2}}\Big)^{4/45}\Big] \\ &+ C_{B,d}(1+M^{1/2})^{4/5}V^{1/5}\Big[\log\Big(2+\frac{Vn^{1/2}}{1+M^{1/2}}\Big)\Big]^{4(s-1)/5}\Big]\cdot n^{-2/5}. \end{aligned}$$

By using Theorem 24 with the function class $\mathcal{F}_{\mathrm{tc}}(V,M)$, we can show that

$$\mathbb{P}\big(\|\hat{f}_{n,V}^{d,s}-f^*\|_{p_0,2}>t\big) \le \epsilon + C\cdot\frac{1+M^3}{t^3}\cdot t_n^3 + C\cdot\frac{\|\xi_1\|_2^3+M^3}{n^{3/2}t^3} + C\cdot\frac{M^3(\|\xi_1\|_5^3 n^{3/5}+M^3)}{n^3t^6}$$

for sufficiently large n and $t\ge t_n$. As a result, again, by repeating the arguments in the proof of Theorem 5, we can find $K>0$ for which

$$\mathbb{P}\big(\|\hat{f}_{n,V}^{d,s}-f^*\|_{p_0,2} > Kn^{-2/5}(\log n)^{4(s-1)/5}\big) < 2\epsilon$$

for sufficiently large n. This proves that

$$\|\hat{f}_{n,V}^{d,s}-f^*\|_{p_0,2}^2 = O_p\big(n^{-4/5}(\log n)^{8(s-1)/5}\big).$$

□

D.8 Proof of Theorem 13

We use the following variant of Theorem 24 in Appendix C to prove Theorem 13. The main differences are that the errors are now assumed to be sub-exponential, and the term $\|\xi_1\|_q^3 n^{3/q}$ in the probability bound is replaced by $K^3[\log(2 + n\sigma^2/K^2)]^3$ as a result of that stronger assumption. Below, we describe the modifications needed in the proof of Theorem 24 to establish Theorem 31.

Theorem 31. *Suppose we are given $(\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(n)}, y_n)$ generated according to the model*

$$y_i = f^*(\mathbf{x}^{(i)}) + \xi_i,$$

where $\mathbf{x}^{(i)}$ are i.i.d. random variables with law P on $[0,1]^d$, and ξ_i are i.i.d. mean-zero sub-exponential errors, independent of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ and satisfying (4.26) for some $K, \sigma > 0$. The function $f^* : [0,1]^d \to \mathbb{R}$ is an unknown regression function to be estimated. Let $\mathcal{F}$ denote a collection of continuous real-valued functions on $[0,1]^d$, and assume $f^* \in \mathcal{F}$. Also, assume that $\|f - f^*\|_\infty \le M$ for all $f \in \mathcal{F}$. Let $\hat{f}$ be an estimator of f^* satisfying

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n \left(y_i - f(\mathbf{x}^{(i)})\right)^2 \right\}$$

with probability at least $1 - \epsilon$ for some $\epsilon \ge 0$.

Suppose there exists $t_n > 0$ such that the following inequalities hold for every $r \ge 1$:

$$\mathbb{E}\left[\sup_{\substack{f \in \mathcal{F} - \{f^*\} \\ \|f\|_{P,2} \le r t_n}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)}) \right| \right] \le r\sqrt{n} t_n^2 \quad \text{and} \quad \mathbb{E}\left[\sup_{\substack{f \in \mathcal{F} - \{f^*\} \\ \|f\|_{P,2} \le r t_n}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)}) \right| \right] \le r\sqrt{n} t_n^2.$$

Here, $\|\cdot\|_{P,2}$ is the norm defined by

$$\|f\|_{P,2} = \left(\mathbb{E}_{\mathbf{X} \sim P}[f^2(\mathbf{X})]\right)^{1/2},$$

and ϵ_i are i.i.d. Rademacher variables independent of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$. Under these assumptions, there exists a universal positive constant C such that, for every $t \ge t_n$,

$$\begin{aligned} \mathbb{P}(\|\hat{f} - f^*\|_{P,2} > t) \le \epsilon + C \cdot \frac{1 + M^3}{t^3} \cdot t_n^3 + C \cdot \frac{\|\xi_1\|_2^3 + M^3}{n^{3/2} t^3} \\ + C \cdot \frac{M^3 K^3 (\log(2 + n\sigma^2/K^2))^3 + M^6}{n^3 t^6}. \end{aligned}$$

Remark 16. *In the proof of Theorem 24, the term $\|\xi_1\|_q^3 n^{3/q}$ in the probability bound arises as an upper bound (up to a constant) on $\mathbb{E}[\max_i |\xi_i|^3]$ under the moment assumption $\|\xi_i\|_q < \infty$ for some $q \ge 3$. Under the stronger assumption that ξ_i are sub-exponential random variables satisfying (4.26) for some $K, \sigma > 0$, we obtain a sharper bound for $\mathbb{E}[\max_i |\xi_i|^3]$ as follows.*

Define $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ by

$$\psi(x) = \exp(x^{1/3}) - 1 - x^{1/3} - \frac{1}{2} \cdot x^{2/3}.$$

It is straightforward to check that ψ is strictly increasing, convex, and invertible. Hence,

$$\begin{aligned}
\mathbb{E}\Big[\max_i |\xi_i|^3\Big] &= K^3 \mathbb{E}\Big[\max_i \Big|\frac{\xi_i}{K}\Big|^3\Big] = K^3 \mathbb{E}\Big[\psi^{-1}\Big(\max_i \psi\Big(\Big|\frac{\xi_i}{K}\Big|^3\Big)\Big)\Big] \\
&\leq K^3 \psi^{-1}\Big(\mathbb{E}\Big[\max_i \psi\Big(\Big|\frac{\xi_i}{K}\Big|^3\Big)\Big]\Big) \leq K^3 \psi^{-1}\Big(\sum_{i=1}^n \mathbb{E}\Big[\psi\Big(\Big|\frac{\xi_i}{K}\Big|^3\Big)\Big]\Big) \\
&= K^3 \psi^{-1}\Big(n\mathbb{E}\Big[\psi\Big(\Big|\frac{\xi_1}{K}\Big|^3\Big)\Big]\Big) \leq K^3 \psi^{-1}\Big(\frac{n\sigma^2}{2K^2}\Big),
\end{aligned}$$

where the first inequality uses the concavity of ψ^{-1}, and the last inequality follows from

$$\mathbb{E}\Big[\psi\Big(\Big|\frac{\xi_1}{K}\Big|^3\Big)\Big] \leq \mathbb{E}\Big[\exp\Big(\Big|\frac{\xi_1}{K}\Big|\Big) - 1 - \Big|\frac{\xi_1}{K}\Big|\Big] \leq \frac{\sigma^2}{2K^2}.$$

Let $C > 0$ be a constant such that

$$\psi(x) \geq \frac{1}{2} \cdot \exp(x^{1/3}) \quad \text{for } x \geq C.$$

Then, for $x \geq \psi(C)$, we have

$$\psi^{-1}(x) \leq [\log(2x)]^3.$$

Consequently,

$$\psi^{-1}\Big(\frac{n\sigma^2}{2K^2}\Big) \leq C + \Big[\log\Big(1 + \frac{n\sigma^2}{K^2}\Big)\Big]^3 \leq C\Big[\log\Big(2 + \frac{n\sigma^2}{K^2}\Big)\Big]^3,$$

from which it follows that

$$\mathbb{E}\Big[\max_i |\xi_i|^3\Big] \leq CK^3\Big[\log\Big(2 + \frac{n\sigma^2}{K^2}\Big)\Big]^3.$$

By using this sharper bound for $\mathbb{E}\big[\max_i |\xi_i|^3\big]$, we can replace the term $\|\xi_1\|_q^3 n^{3/q}$ in the probability bound in Theorem 24 with $K^3[\log(2 + n\sigma^2/K^2)]^3$ and establish Theorem 31.

Proof of Theorem 13. Fix $\epsilon_0 > 0$. By the same argument as in the proof of Theorem 12, there exists $M > 0$ such that

$$\mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f^*\|_\infty \leq M\big) \geq 1 - \epsilon_0$$

for sufficiently large n. Define

$$\mathcal{F}_{\mathrm{tc}}(V, M) = \big\{f \in \mathcal{F}_{\mathrm{tc}}^{d,s} : V_{\mathrm{mars}}(f) \leq V \text{ and } \|f - f^*\|_\infty \leq M\big\}$$

and

$$B_{p_0}(M,t) = \big\{f \in \mathcal{F}_{\mathrm{tc}}^{d,s} : \|f - f^*\|_\infty \le M \text{ and } \|f - f^*\|_{p_0,2} \le t\big\}.$$

Then, $f^* \in \mathcal{F}_{\mathrm{tc}}(V,M)$, and for sufficiently large n, we have

$$\hat{f}_{n,V}^{d,s} \in \operatorname*{argmin}_f \bigg\{ \sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)})\big)^2 : f \in \mathcal{F}_{\mathrm{tc}}(V,M) \bigg\}$$

with probability at least $1 - \epsilon_0$.

We first consider the special case $f^* = 0$. In this case,

$$\mathcal{F}_{\mathrm{tc}}(V,M) = \big\{f \in \mathcal{F}_{\mathrm{tc}}^{d,s} : V_{\mathrm{mars}}(f) \le V \text{ and } \|f\|_\infty \le M\big\}$$

and

$$B_{p_0}(M,t) = \big\{f \in \mathcal{F}_{\mathrm{tc}}^{d,s} : \|f\|_\infty \le M \text{ and } \|f\|_{p_0,2} \le t\big\}.$$

Our goal is to apply Theorem 31. To this end, we need to bound

$$\mathbb{E}\bigg[\sup_{\substack{f \in \mathcal{F}_{\mathrm{tc}}(V,M) \\ \|f\|_{p_0,2} \le t}} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\bigg]. \tag{D.31}$$

Since the Rademacher variables ϵ_i are also sub-exponential, the same bound applies to

$$\mathbb{E}\bigg[\sup_{\substack{f \in \mathcal{F}_{\mathrm{tc}}(V,M) \\ \|f\|_{p_0,2} \le t}} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\bigg].$$

We first enlarge the function class over which the supremum is taken:

$$\mathbb{E}\bigg[\sup_{\substack{f \in \mathcal{F}_{\mathrm{tc}}(V,M) \\ \|f\|_{p_0,2} \le t}} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\bigg] \le \mathbb{E}\bigg[\sup_{f \in B_{p_0}(M,t)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\bigg]. \tag{D.32}$$

By the same argument as in the proof of Lemma 16 in Appendix C, the expectation on the right-hand side of (D.32) can be bounded in terms of the bracketing entropy integral:

$$\begin{aligned}
&\mathbb{E}\bigg[\sup_{f \in B_{p_0}(M,t)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\bigg] \\
&\qquad \le C J_{[\,]}(t, B_{p_0}(M,t), \|\cdot\|_{p_0,2}) \cdot \bigg(\sigma + KM \cdot \frac{J_{[\,]}(t, B_{p_0}(M,t), \|\cdot\|_{p_0,2})}{t^2\sqrt{n}}\bigg),
\end{aligned} \tag{D.33}$$

where

$$J_{[\,]}(t, B_{p_0}(M,t), \|\cdot\|_{p_0,2}) = \int_0^t \sqrt{1 + \log N_{[\,]}(\epsilon, B_{p_0}(M,t), \|\cdot\|_{p_0,2})}\, d\epsilon,$$

and $N_{[\,]}(\epsilon, \mathcal{G}, \|\cdot\|)$ denotes the ϵ-bracketing number of $\mathcal{G}$ with respect to $\|\cdot\|$. Since $b \leq p_0(\mathbf{x}) \leq B$ for all $\mathbf{x} \in [0,1]^d$, we also have

$$N_{[\,]}(\epsilon, B_{p_0}(M,t), \|\cdot\|_{p_0,2}) \leq N_{[\,]}(\epsilon/B^{1/2}, B(M, t/b^{1/2}), \|\cdot\|_2), \tag{D.34}$$

where

$$B(M,t) = \big\{f \in \mathcal{F}_{\text{tc}}^{d,s} : \|f\|_\infty \leq M \text{ and } \|f\|_2 \leq t\big\}.$$

For $\boldsymbol{\delta} \in \{0,1\}^d$, define

$$P^{(\boldsymbol{\delta})} = \prod_{j=1}^d P_{\delta_j},$$

where $P_0 = [0, 1/2]$ and $P_1 = [1/2, 1]$. Also, let $B^{(\boldsymbol{\delta})}(M,t)$ denote the collection of functions $g : P^{(\boldsymbol{\delta})} \to \mathbb{R}$ such that $g = f|_{P^{(\boldsymbol{\delta})}}$ for some $f \in B(M,t)$. It then follows that

$$\log N_{[\,]}(\epsilon, B(M,t), \|\cdot\|_2) \leq \sum_{\boldsymbol{\delta} \in \{0,1\}^d} \log N_{[\,]}\Big(\frac{\epsilon}{2^{d/2}}, B^{(\boldsymbol{\delta})}(M,t), \|\cdot\|_2\Big). \tag{D.35}$$

From this point forward, we focus on bounding $\log N_{[\,]}(\epsilon, B^{(\mathbf{0})}(M,t), \|\cdot\|_2)$. An entirely analogous argument applies to $\log N_{[\,]}(\epsilon, B^{(\boldsymbol{\delta})}(M,t), \|\cdot\|_2)$ for $\boldsymbol{\delta} \in \{0,1\}^d \setminus \{\mathbf{0}\}$. Fix $\epsilon > 0$, and define

$$R = \max\left(\left\lceil \log\Big(\frac{dM^2}{\epsilon^2}\Big)/\log 2\right\rceil - (d-3), 1\right).$$

This choice of R ensures that

$$\frac{M^2}{2^{d-2}} \cdot \frac{d}{2^R} \leq \frac{\epsilon^2}{2}.$$

Now, consider the set

$$P_{\text{int}}^{(\mathbf{0})} = \Big[\frac{1}{2^{R+1}}, \frac{1}{2}\Big]^d,$$

and let $B_{\text{int}}^{(\mathbf{0})}(t)$ denote the collection of functions $g : P_{\text{int}}^{(\mathbf{0})} \to \mathbb{R}$ such that $g = f|_{P_{\text{int}}^{(\mathbf{0})}}$ for some $f \in \mathcal{F}_{\text{tc}}^{d,s}$ with $\|f\|_2 \leq t$. Intuitively, $P_{\text{int}}^{(\mathbf{0})}$ can be thought of as an interior of $P^{(\mathbf{0})} = [0,1/2]^d$ obtained by peeling off its boundary. We claim that

$$\log N_{[\,]}\big(\epsilon, B^{(\mathbf{0})}(M,t), \|\cdot\|_2\big) \leq \log N_{[\,]}\Big(\frac{\epsilon}{\sqrt{2}}, B_{\text{int}}^{(\mathbf{0})}(t), \|\cdot\|_2\Big). \tag{D.36}$$

To prove this, let $g \in B^{(\mathbf{0})}(M,t)$. By definition, $g = f|_{P^{(\mathbf{0})}}$ for some $f \in B(M,t)$, and hence, $g|_{P_{\text{int}}^{(\mathbf{0})}} \in B_{\text{int}}^{(\mathbf{0})}(t)$. Let $[h_1, h_2]$ be a bracket containing $g|_{P_{\text{int}}^{(\mathbf{0})}}$ with $\|h_1 - h_2\|_2 \leq \epsilon/\sqrt{2}$. Consider extensions h_1^{ext} and h_2^{ext} of h_1 and h_2 to $P^{(\mathbf{0})}$ for which $h_1^{\text{ext}}(\mathbf{x}) = -M$ for $\mathbf{x} \notin P_{\text{int}}^{(\mathbf{0})}$ and $h_2^{\text{ext}}(\mathbf{x}) = M$ for $\mathbf{x} \notin P_{\text{int}}^{(\mathbf{0})}$. Then, $g \in [h_1^{\text{ext}}, h_2^{\text{ext}}]$ and

$$\|h_1^{\text{ext}} - h_2^{\text{ext}}\|_2^2 \leq \|h_1 - h_2\|_2^2 + 4M^2 \cdot \left(\frac{1}{2^d} - \Big(\frac{1}{2} - \frac{1}{2^{R+1}}\Big)^d\right)$$

$$\le \|h_1 - h_2\|_2^2 + \frac{M^2}{2^{d-2}} \cdot \Big(1 - \Big(1 - \frac{1}{2^R}\Big)^d\Big)$$
$$\le \|h_1 - h_2\|_2^2 + \frac{M^2}{2^{d-2}} \cdot \frac{d}{2^R} \le \|h_1 - h_2\|_2^2 + \frac{\epsilon^2}{2} \le \epsilon^2.$$

This proves the claim (D.36).

Next, we split $P_{\text{int}}^{(\mathbf{0})}$ into dyadic pieces. Let $\mathcal{R} = \{1, \ldots, R\}^d$, and for each $\mathbf{r} = (r_1, \ldots, r_d) \in \mathcal{R}$, define

$$Q_{\mathbf{r}} = \prod_{j=1}^{d} Q_{r_j},$$

where $Q_r = [1/2^{r+1}, 1/2^r]$ for $r = 1, \ldots, R$. Also, for each $\mathbf{r} \in \mathcal{R}$, let $B_{\mathbf{r}}^{(\mathbf{0})}(t)$ denote the collection of functions $g : Q_{\mathbf{r}} \to \mathbb{R}$ such that $g = f|_{Q_{\mathbf{r}}}$ for some $f \in \mathcal{F}_{\text{tc}}^{d,s}$ with $\|f\|_2 \le t$. Since

$$P_{\text{int}}^{(\mathbf{0})} = \bigcup_{\mathbf{r} \in \mathcal{R}} Q_{\mathbf{r}},$$

we have

$$\log N_{[\,]}\Big(\frac{\epsilon}{\sqrt{2}}, B_{\text{int}}^{(\mathbf{0})}(t), \|\cdot\|_2\Big) \le \sum_{\mathbf{r} \in \mathcal{R}} \log N_{[\,]}\Big(\frac{\epsilon}{\sqrt{2|\mathcal{R}|}}, B_{\mathbf{r}}^{(\mathbf{0})}(t), \|\cdot\|_2\Big). \tag{D.37}$$

We now apply the following lemma, proved in Appendix D.14, to bound each term on the right-hand side of (D.37).

Lemma 33. *Assume that $f \in \mathcal{F}_{tc}^{d,s}$ and $\|f\|_2 \le t$. Then, for every $\emptyset \neq S \subseteq [d]$, $1/2^{r_j+1} \le s_j < t_j \le 1/2^{r_j}$ for $j \in S$, and $t_j \in [1/2^{r_j+1}, 1/2^{r_j}]$ for $j \notin S$, we have*

$$\begin{bmatrix} s_j, t_j, & j \in S \\ t_j, & j \notin S \end{bmatrix} ; \ f \Bigg] \ge -C_d t \cdot 2^{r_+/2} \cdot 2^{\sum_{j \in S} r_j} \tag{D.38}$$

if $\mathbf{r} = (r_1, \ldots, r_d) \in \mathbb{N}^d$, and

$$\begin{bmatrix} s_j, t_j, & j \in S \\ t_j, & j \notin S \end{bmatrix} ; \ f \Bigg] \le C_d t \cdot 2^{r_+/2} \cdot 2^{\sum_{j \in S} r_j} \tag{D.39}$$

if $\mathbf{r} = (r_1, \ldots, r_d) \in \mathbb{Z}_{\ge 0}^d$, where $r_+ = \sum_{j=1}^d r_j$.

Fix $\mathbf{r} = (r_1, \ldots, r_d) \in \mathcal{R}$, and suppose $g \in B_{\mathbf{r}}^{(\mathbf{0})}(t)$. Define $h : [0,1]^d \to \mathbb{R}$ by

$$h(x_1, \ldots, x_d) = g\Big(\frac{x_1 + 1}{2^{r_1+1}}, \ldots, \frac{x_d + 1}{2^{r_d+1}}\Big),$$

so that h is a domain-scaled version of g. By the definition of $B_{\mathbf{r}}^{(\mathbf{0})}(t)$, there exists $f \in \mathcal{F}_{\text{tc}}^{d,s}$ with $\|f\|_2 \le t$ such that $g = f|_{Q_{\mathbf{r}}}$. Since $f \in \mathcal{F}_{\text{tc}}^{d,s}$, it follows that $h \in \mathcal{F}_{\text{tcp}}^{d}$, and h satisfies the interaction restriction condition (4.16) for every subset S of $[d]$ with $|S| > s$. By Lemma 33, for every nonempty subset S of $[d]$,

$$\begin{aligned}\lim_{u \to 0+} \begin{bmatrix} 0, u, & j \in S & ; & h \\ 0, & j \notin S & & \end{bmatrix} &= \prod_{j \in S} \frac{1}{2^{r_j+1}} \cdot \lim_{u \to 0+} \begin{bmatrix} 1/2^{r_j+1}, (1+u)/2^{r_j+1}, & j \in S & ; & f \\ 1/2^{r_j+1}, & j \notin S & & \end{bmatrix} \\ &\le \frac{1}{2^{|S|}} \cdot C_d t \cdot 2^{r_+/2}\end{aligned}$$

and

$$\begin{aligned}\lim_{u \to 1-} \begin{bmatrix} u, 1, & j \in S & ; & h \\ 0, & j \notin S & & \end{bmatrix} &= \prod_{j \in S} \frac{1}{2^{r_j+1}} \cdot \lim_{u \to 1-} \begin{bmatrix} (1+u)/2^{r_j+1}, 1/2^{r_j}, & j \in S & ; & f \\ 1/2^{r_j+1}, & j \notin S & & \end{bmatrix} \\ &\ge -\frac{1}{2^{|S|}} \cdot C_d t \cdot 2^{r_+/2}.\end{aligned}$$

Hence, Proposition 5 ensures $h \in \mathcal{F}^{d,s}_{\mathrm{tc}}$. Also, $V_{\mathrm{mars}}(h)$ of h can be bounded as

$$\begin{aligned}V_{\mathrm{mars}}(h) &= \sum_{\emptyset \neq S \subseteq [d]} \left(\lim_{u \to 0+} \begin{bmatrix} 0, u, & j \in S & ; & h \\ 0, & j \notin S & & \end{bmatrix} - \lim_{u \to 1-} \begin{bmatrix} u, 1, & j \in S & ; & h \\ 0, & j \notin S & & \end{bmatrix} \right) \\ &\le \sum_{\emptyset \neq S \subseteq [d]} \frac{1}{2^{|S|}} \cdot C_d t \cdot 2^{r_+/2} \le C_d t \cdot 2^{r_+/2}.\end{aligned}$$

Moreover,

$$\|g\|_2^2 = \int_{Q_{\mathbf{r}}} g^2(\mathbf{x})\, d\mathbf{x} = \prod_{j=1}^{d} \frac{1}{2^{r_j+1}} \cdot \int_{[0,1]^d} h^2(\mathbf{x})\, d\mathbf{x} = \frac{1}{2^{r_+ + d}} \cdot \|h\|_2^2.$$

Consequently, it follows that

$$\log N_{[\,]}\Big(\frac{\epsilon}{\sqrt{2|\mathcal{R}|}}, B^{(\mathbf{0})}_{\mathbf{r}}(t), \|\cdot\|_2\Big) \le \log N_{[\,]}\Big(\frac{2^{(r_+ + d - 1)/2}\epsilon}{|\mathcal{R}|^{1/2}}, D(C_d t \cdot 2^{r_+/2}, t \cdot 2^{(r_+ + d)/2}), \|\cdot\|_2\Big),$$

where

$$D(V, t) := \big\{ f \in \mathcal{F}^{d,s}_{\mathrm{tc}} : V_{\mathrm{mars}}(f) \le V \text{ and } \|f\|_2 \le t \big\}.$$

The following lemma, proved in Appendix D.15, provides an upper bound on the bracketing number of $D(V, t)$ for $V, t > 0$.

Lemma 34. *There exists a positive constant C_d depending on d such that*

$$\log N_{[\,]}(\epsilon, D(V,t), \|\cdot\|_2) \le C_d \log\Big(2 + \frac{t+V}{\epsilon}\Big) + C_d \Big(2 + \frac{V}{\epsilon}\Big)^{1/2} \Big[\log\Big(2 + \frac{V}{\epsilon}\Big)\Big]^{2(s-1)}$$

for every $V, t > 0$ and $\epsilon > 0$.

Applying Lemma 34, we obtain

$$\begin{aligned}&\log N_{[\,]}\Big(\frac{\epsilon}{\sqrt{2|\mathcal{R}|}}, B^{(\mathbf{0})}_{\mathbf{r}}(t), \|\cdot\|_2\Big)\\ &\qquad \le C_d \log\Big(2 + \frac{C_d t|\mathcal{R}|^{1/2}}{\epsilon}\Big) + C_d\Big(2 + \frac{C_d t|\mathcal{R}|^{1/2}}{\epsilon}\Big)^{1/2}\Big[\log\Big(2 + \frac{C_d t|\mathcal{R}|^{1/2}}{\epsilon}\Big)\Big]^{2(s-1)}.\end{aligned}$$

Combining this with (D.36) and (D.37), we deduce

$$\begin{aligned}\log N_{[\,]}\big(\epsilon, B^{(\mathbf{0})}(M,t), \|\cdot\|_2\big) &\le \log N_{[\,]}\Big(\frac{\epsilon}{\sqrt{2}}, B^{(\mathbf{0})}_{\mathrm{int}}(t), \|\cdot\|_2\Big)\\ &\le \sum_{\mathbf{r}\in\mathcal{R}} \log N_{[\,]}\Big(\frac{\epsilon}{\sqrt{2|\mathcal{R}|}}, B^{(\mathbf{0})}_{\mathbf{r}}(t), \|\cdot\|_2\Big)\\ &\le C_d|\mathcal{R}|\cdot\log\Big(2 + \frac{C_d t|\mathcal{R}|^{1/2}}{\epsilon}\Big)\\ &\qquad + C_d|\mathcal{R}|\Big(2 + \frac{C_d t|\mathcal{R}|^{1/2}}{\epsilon}\Big)^{1/2}\Big[\log\Big(2 + \frac{C_d t|\mathcal{R}|^{1/2}}{\epsilon}\Big)\Big]^{2(s-1)}.\end{aligned}$$

Next, by (D.35),

$$\begin{aligned}&\log N_{[\,]}(\epsilon, B(M,t), \|\cdot\|_2) \le \sum_{\boldsymbol{\delta}\in\{0,1\}^d} \log N_{[\,]}\Big(\frac{\epsilon}{2^{d/2}}, B^{(\boldsymbol{\delta})}(M,t), \|\cdot\|_2\Big)\\ &\quad\le C_d|\mathcal{R}|\cdot\log\Big(2 + \frac{C_d t|\mathcal{R}|^{1/2}}{\epsilon}\Big) + C_d|\mathcal{R}|\Big(2 + \frac{C_d t|\mathcal{R}|^{1/2}}{\epsilon}\Big)^{1/2}\Big[\log\Big(2 + \frac{C_d t|\mathcal{R}|^{1/2}}{\epsilon}\Big)\Big]^{2(s-1)}\\ &\quad\le C_{d,M}\log\Big(2 + \frac{C_{d,M}t}{\epsilon}\cdot\Big(\log\Big(2+\frac{1}{\epsilon}\Big)\Big)^{d/2}\Big)\cdot\Big[\log\Big(2+\frac{1}{\epsilon}\Big)\Big]^d\\ &\qquad + C_{d,M}\Big(2 + \frac{C_{d,M}t}{\epsilon}\cdot\Big(\log\Big(2+\frac{1}{\epsilon}\Big)\Big)^{d/2}\Big)^{1/2}\\ &\qquad\qquad \cdot\Big[\log\Big(2+\frac{1}{\epsilon}\Big)\Big]^d\Big[\log\Big(2 + \frac{C_{d,M}t}{\epsilon}\cdot\Big(\log\Big(2+\frac{1}{\epsilon}\Big)\Big)^{d/2}\Big)\Big]^{2(s-1)}\\ &\quad\le C_{d,M}\log\Big(3+\frac{t}{\epsilon}\Big)\cdot\Big[\log\Big(3+\frac{1}{\epsilon}\Big)\Big]^d + C_{d,M}\log\log\Big(3+\frac{1}{\epsilon}\Big)\cdot\Big[\log\Big(3+\frac{1}{\epsilon}\Big)\Big]^d\\ &\qquad + C_{d,M}\Big(3+\frac{t}{\epsilon}\Big)^{1/2}\Big[\log\Big(3+\frac{1}{\epsilon}\Big)\Big]^{5d/4}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{2(s-1)}\\ &\qquad + C_{d,M}\Big(3+\frac{t}{\epsilon}\Big)^{1/2}\Big[\log\Big(3+\frac{1}{\epsilon}\Big)\Big]^{5d/4}\Big[\log\log\Big(3+\frac{1}{\epsilon}\Big)\Big]^{2(s-1)},\end{aligned}$$

where the third inequality uses

$$|\mathcal{R}| = R^d \le C_{d,M}\Big[\log\Big(2+\frac{1}{\epsilon}\Big)\Big]^d,$$

which follows from

$$R \le \max\Big(\log\Big(\frac{dM^2}{\epsilon^2}\Big)/\log 2 - (d-2), 1\Big) \le C_{d,M}\log\Big(2+\frac{1}{\epsilon}\Big).$$

If $t \geq 1$, then

$$\log\Big(3+\frac{1}{\epsilon}\Big) \leq \log\Big(3+\frac{t}{\epsilon}\Big).$$

If instead $t < 1$, then

$$\log\Big(3+\frac{1}{\epsilon}\Big) = \log\Big(3t+\frac{t}{\epsilon}\Big) - \log t \leq \log\Big(3+\frac{t}{\epsilon}\Big) + \log\frac{1}{t} \leq \log\Big(3+\frac{t}{\epsilon}\Big) + \log\Big(3+\frac{1}{t}\Big).$$

Thus, in all cases,

$$\log\Big(3+\frac{1}{\epsilon}\Big) \leq \log\Big(3+\frac{t}{\epsilon}\Big) + \log\Big(3+\frac{1}{t}\Big).$$

As a result,

$$\begin{aligned}
\log N_{[\,]}(\epsilon, B(M,t), \|\cdot\|_2) \leq\; & C_{d,M}\log\Big(3+\frac{t}{\epsilon}\Big)\cdot\Big[\log\Big(3+\frac{t}{\epsilon}\Big)+\log\Big(3+\frac{1}{t}\Big)\Big]^{d+1/e} \\
& + C_{d,M}\Big(3+\frac{t}{\epsilon}\Big)^{1/2}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)+\log\Big(3+\frac{1}{t}\Big)\Big]^{5d/4}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{2(s-1)} \\
& + C_{d,M}\Big(3+\frac{t}{\epsilon}\Big)^{1/2}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)+\log\Big(3+\frac{1}{t}\Big)\Big]^{5d/4+2(s-1)/e} \\
\leq\; & C_{d,M}\log\Big(3+\frac{t}{\epsilon}\Big)\cdot\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{d+1/e} + C_{d,M}\log\Big(3+\frac{t}{\epsilon}\Big)\cdot\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{d+1/e} \\
& + C_{d,M}\Big(3+\frac{t}{\epsilon}\Big)^{1/2}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{5d/4+2(s-1)} \\
& + C_{d,M}\Big(3+\frac{t}{\epsilon}\Big)^{1/2}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{2(s-1)}\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{5d/4+2(s-1)/e}.
\end{aligned} \tag{D.40}$$

For the first inequality, we used $\log\log x \leq (\log x)^{1/e}$ for all $x > 1$. By (D.34), it follows that

$$\begin{aligned}
\log N_{[\,]}(\epsilon, B_{p_0}(M,t), \|\cdot\|_{p_0,2}) \leq\; & C_{b,B,d,M}\log\Big(3+\frac{t}{\epsilon}\Big)\cdot\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{d+1/e} \\
& + C_{b,B,d,M}\log\Big(3+\frac{t}{\epsilon}\Big)\cdot\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{d+1/e} + C_{b,B,d,M}\Big(3+\frac{t}{\epsilon}\Big)^{1/2}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{5d/4+2(s-1)} \\
& + C_{b,B,d,M}\Big(3+\frac{t}{\epsilon}\Big)^{1/2}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{2(s-1)}\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{5d/4+2(s-1)/e}.
\end{aligned}$$

Next, we bound the integrals appearing in the bracketing entropy integral. Observe that

$$\int_0^t \Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{1/2}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{d/2+1/2e} d\epsilon \leq Ct,$$

$$\int_0^t \Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{1/2}\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{d/2+1/2e} d\epsilon \leq Ct\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{d/2+1/2e},$$

$$\int_0^t \Big(3+\frac{t}{\epsilon}\Big)^{1/4}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{5d/8+s-1} d\epsilon \leq Ct,$$

and

$$\int_0^t \Big(3+\frac{t}{\epsilon}\Big)^{1/4}\Big[\log\Big(3+\frac{t}{\epsilon}\Big)\Big]^{s-1}\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{5d/8+(s-1)/e} d\epsilon \le Ct\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{5d/8+(s-1)/e}.$$

Hence, the bracketing entropy integral satisfies

$$\begin{aligned} J_{[\,]}(t, B_{p_0}(M,t), \|\cdot\|_{p_0,2}) &= \int_0^t \sqrt{1+\log N_{[\,]}(\epsilon, B_{p_0}(M,t), \|\cdot\|_{p_0,2})}\, d\epsilon \\ &\le C_{b,B,d,M}\, t\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{a(d,s)}, \end{aligned}$$

where

$$a(d,s) := \max\Big(\frac{5d}{8}+\frac{s-1}{e}, \frac{d}{2}+\frac{1}{2e}\Big).$$

Combining this with (D.32) and (D.33) gives

$$\begin{aligned} \mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_{\mathrm{tc}}(V,M)\\ \|f\|_{p_0,2}\le t}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] &\le \mathbb{E}\Big[\sup_{f\in B_{p_0}(M,t)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] \\ &\le CJ_{[\,]}(t, B_{p_0}(M,t), \|\cdot\|_{p_0,2})\cdot\Big(\sigma + KM\cdot\frac{J_{[\,]}(t, B_{p_0}(M,t), \|\cdot\|_{p_0,2})}{t^2\sqrt{n}}\Big) \\ &\le C_{b,B,d,M}\sigma t\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{a(d,s)} + C_{b,B,d,M}Kn^{-1/2}\Big[\log\Big(3+\frac{1}{t}\Big)\Big]^{2a(d,s)} =: \Psi(t). \end{aligned}$$

Let

$$t_n = C_{b,B,d,M}(1+\sigma+K^{1/2})\cdot n^{-1/2}\big(\log(3+n^{1/2})\big)^{a(d,s)}.$$

Then, we have

$$\Psi(t_n) \le \sqrt{n}t_n^2.$$

Since $t\mapsto\Psi(t)/t$ is decreasing, for every $r\ge 1$,

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_{\mathrm{tc}}(V,M)\\ \|f\|_{p_0,2}\le rt_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] \le \Psi(rt_n) \le r\Psi(t_n) \le r\sqrt{n}t_n^2.$$

Repeating all preceding arguments with the Rademacher variables ϵ_i in place of ξ_i also yields

$$\mathbb{E}\Big[\sup_{\substack{f\in\mathcal{F}_{\mathrm{tc}}(V,M)\\ \|f\|_{p_0,2}\le rt_n}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] \le r\sqrt{n}t_n^2 \quad \text{for every } r\ge 1.$$

Applying Theorem 31 therefore gives

$\mathbb{P}\big(\|\hat{f}^{d,s}_{n,V} - f^*\|_{p_0,2} > t\big)$

$$\le \epsilon_0 + C \cdot \frac{1+M^3}{t^3} \cdot t_n^3 + C \cdot \frac{\|\xi_1\|_2^3 + M^3}{n^{3/2}t^3} + C \cdot \frac{M^3K^3[\log(2+n\sigma^2/K^2)]^3 + M^6}{n^3t^6}.$$

It follows that for each $L > 0$,

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f^*\|_{p_0,2} > Ln^{-1/2}(\log n)^{a(d,s)}\big) \\
&\quad \le \epsilon_0 + \limsup_{n\to\infty}\Bigg[C \cdot \frac{(1+M^3)t_n^3}{L^3 n^{-3/2}(\log n)^{3a(d,s)}} + C \cdot \frac{\|\xi_1\|_2^3 + M^3}{L^3(\log n)^{3a(d,s)}} \\
&\qquad\qquad + C \cdot \frac{M^3K^3[\log(2+n\sigma^2/K^2)]^3 + M^6}{L^6(\log n)^{6a(d,s)}}\Bigg] \\
&\quad \le \epsilon_0 + C_{b,B,d,M} \cdot \frac{(1+\sigma+K^{1/2})^3}{L^3},
\end{aligned}$$

where the second inequality uses that $6a(d,s) > 3d \ge 3$. Hence, for sufficiently large L,

$$\limsup_{n\to\infty} \mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f^*\|_{p_0,2} > Ln^{-1/2}(\log n)^{a(d,s)}\big) < 2\epsilon_0,$$

which proves that

$$\|\hat{f}_{n,V}^{d,s} - f^*\|_{p_0,2} = O_p\big(n^{-1/2}(\log n)^{a(d,s)}\big).$$

Note that the constants underlying O_p depend on b, B, d, V, K, and σ, with the dependence on V entering through M.

We now prove the theorem for the general case. Suppose that f^* is multi-affine of the form (4.23) on each rectangle

$$R_{\mathbf{l}} := \prod_{j=1}^{d}\big[v_{l_j-1}^{(j)}, v_{l_j}^{(j)}\big], \quad \mathbf{l} = (l_1, \dots, l_d) \in \prod_{j=1}^{d}[m_j],$$

where $0 = v_0^{(j)} < \cdots < v_{m_j}^{(j)} = 1$ for $j \in [d]$. Also, assume that $\prod_{j=1}^{d} m_j = m(f^*) =: m$. As in the special case $f^* = 0$, our goal is to apply Theorem 31. To this end, we aim to bound

$$\mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}_{\mathrm{tc}}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le t}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i f(\mathbf{x}^{(i)})\Big|\Bigg].$$

As before, we first enlarge the function class:

$$\mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}_{\mathrm{tc}}(V,M)-\{f^*\}\\ \|f\|_{p_0,2}\le t}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i f(\mathbf{x}^{(i)})\Big|\Bigg] \le \mathbb{E}\Bigg[\sup_{f\in B_{p_0}(M,t)-\{f^*\}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i f(\mathbf{x}^{(i)})\Big|\Bigg].$$

Analogous to (D.33), this expectation can be bounded using the bracketing entropy integral as follows:

$$\mathbb{E}\Bigg[\sup_{f\in B_{p_0}(M,t)-\{f^*\}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i f(\mathbf{x}^{(i)})\Big|\Bigg] \le CJ_{[\,]}(t, B_{p_0}(M,t) - \{f^*\}, \|\cdot\|_{p_0,2})$$

$$\cdot\Big(\sigma + KM \cdot \frac{J_{[\,]}(t, B_{p_0}(M,t) - \{f^*\}, \|\cdot\|_{p_0,2})}{t^2\sqrt{n}}\Big).$$

Since $b \le p_0(\mathbf{x}) \le B$ for all $\mathbf{x} \in [0,1]^d$, we also have

$$N_{[\,]}(\epsilon, B_{p_0}(M,t) - \{f^*\}, \|\cdot\|_{p_0,2}) \le N_{[\,]}(\epsilon/B^{1/2}, \bar{B}(M, t/b^{1/2}) - \{f^*\}, \|\cdot\|_2), \tag{D.41}$$

where

$$\bar{B}(M,t) = \big\{f \in \mathcal{F}_{\mathrm{tc}}^{d,s} : \|f - f^*\|_\infty \le M \text{ and } \|f - f^*\|_2 \le t\big\}.$$

Next, we decompose $[0,1]^d$ into m rectangles $R_{\mathbf{l}}$, on which f^* is multi-affine. For each $\mathbf{l} = (l_1, \dots, l_d) \in \prod_{j=1}^d [m_j]$, let $A_{\mathbf{l}}$ denote the area of $R_{\mathbf{l}}$:

$$A_{\mathbf{l}} = \prod_{j=1}^d \big(v_{l_j}^{(j)} - v_{l_j-1}^{(j)}\big),$$

and let $\bar{B}_{\mathbf{l}}(M,t)$ be the collection of functions $g : R_{\mathbf{l}} \to \mathbb{R}$ such that $g = f|_{R_{\mathbf{l}}}$ for some $f \in \bar{B}(M,t)$. It then follows that

$$\log N_{[\,]}(\epsilon, \bar{B}(M,t) - \{f^*\}, \|\cdot\|_2) \le \sum_{\mathbf{l} \in \prod_{j=1}^d [m_j]} \log N_{[\,]}\big(\epsilon/\sqrt{m}, \bar{B}_{\mathbf{l}}(M,t) - \{f^*\}, \|\cdot\|_2\big). \tag{D.42}$$

Recall that $m = m(f^*) = \prod_{j=1}^d m_j$.

Suppose $g \in \bar{B}_{\mathbf{l}}(M,t)$. By the definition of $\bar{B}_{\mathbf{l}}(M,t)$, there exists $f \in \bar{B}(M,t)$ with $\|f - f^*\|_\infty \le M$ and $\|f - f^*\|_2 \le t$ such that $g = f|_{R_{\mathbf{l}}}$. Define $h : [0,1]^d \to \mathbb{R}$ by

$$h(x_1, \dots, x_d) = (g - f^*)\Big(v_{l_1-1}^{(1)} + \big(v_{l_1}^{(1)} - v_{l_1-1}^{(1)}\big)x_1, \dots, v_{l_d-1}^{(d)} + \big(v_{l_d}^{(d)} - v_{l_d-1}^{(d)}\big)x_d\Big).$$

Since $f, f^* \in \mathcal{F}_{\mathrm{tc}}^{d,s}$ and f^* is multi-affine on $R_{\mathbf{l}}$, it follows that $h \in \mathcal{F}_{\mathrm{tcp}}^d$, and h satisfies the interaction restriction condition (4.16) for every subset S of $[d]$ with $|S| > s$. By induction, it can also be readily verified that for every nonempty subset $S \subseteq [d]$,

$$\begin{aligned}&\lim_{u \to 0+} \begin{bmatrix} 0, u, & j \in S \\ 0, & j \notin S \end{bmatrix} ; \ h \Bigg] \\ &\quad = \prod_{j \in S} \big(v_{l_j}^{(j)} - v_{l_j-1}^{(j)}\big) \cdot \lim_{u \to 0+} \begin{bmatrix} v_{l_j-1}^{(j)}, v_{l_j-1}^{(j)} + (v_{l_j}^{(j)} - v_{l_j-1}^{(j)})u, & j \in S \\ v_{l_j-1}^{(j)}, & j \notin S \end{bmatrix} ; \ f - f^* \Bigg] < \infty\end{aligned}$$

and

$$\begin{aligned}&\lim_{u \to 1-} \begin{bmatrix} u, 1, & j \in S \\ 0, & j \notin S \end{bmatrix} ; \ h \Bigg] \\ &\quad = \prod_{j \in S} \big(v_{l_j}^{(j)} - v_{l_j-1}^{(j)}\big) \cdot \lim_{u \to 1-} \begin{bmatrix} v_{l_j-1}^{(j)} + (v_{l_j}^{(j)} - v_{l_j-1}^{(j)})u, v_{l_j}^{(j)}, & j \in S \\ v_{l_j-1}^{(j)}, & j \notin S \end{bmatrix} ; \ f - f^* \Bigg] > -\infty.\end{aligned}$$

By Proposition 5, these facts ensure that $h \in \mathcal{F}_{\text{tc}}^{d,s}$. Furthermore, since

$$\|g - f^*\|_2^2 = A_{\mathbf{l}} \cdot \|h\|_2^2,$$

we obtain

$$\log N_{[\,]}\big(\epsilon/\sqrt{m}, \bar{B}_{\mathbf{l}}(M,t) - \{f^*\}, \|\cdot\|_2\big) \le \log N_{[\,]}\bigg(\frac{\epsilon}{m^{1/2}A_{\mathbf{l}}^{1/2}}, B\Big(M, \frac{t}{A_{\mathbf{l}}^{1/2}}\Big), \|\cdot\|_2\bigg),$$

where

$$B(M,t) = \{f \in \mathcal{F}_{\text{tc}}^{d,s} : \|f\|_\infty \le M \text{ and } \|f\|_2 \le t\}.$$

Using the bound (D.40) established in the special case $f^* = 0$, we obtain

$$\begin{aligned}
\log N_{[\,]}\big(\epsilon/\sqrt{m}, \bar{B}_{\mathbf{l}}(M,t) - \{f^*\}, \|\cdot\|_2\big) &\le C_{d,M} \log\left(3 + \frac{m^{1/2}t}{\epsilon}\right) \cdot \left[\log\left(3 + \frac{m^{1/2}t}{\epsilon}\right)\right]^{d+1/e} \\
&\quad + C_{d,M} \log\left(3 + \frac{m^{1/2}t}{\epsilon}\right) \cdot \left[\log\left(3 + \frac{A_{\mathbf{l}}^{1/2}}{t}\right)\right]^{d+1/e} + C_{d,M}\left[\log\left(3 + \frac{m^{1/2}t}{\epsilon}\right)\right]^{5d/4+2(s-1)} \\
&\quad + C_{d,M}\left(3 + \frac{m^{1/2}t}{\epsilon}\right)^{1/2}\left[\log\left(3 + \frac{m^{1/2}t}{\epsilon}\right)\right]^{2(s-1)}\left[\log\left(3 + \frac{A_{\mathbf{l}}^{1/2}}{t}\right)\right]^{5d/4+2(s-1)/e} \\
&\le C_{d,M,m} \log\left(3 + \frac{t}{\epsilon}\right) \cdot \left[\log\left(3 + \frac{t}{\epsilon}\right)\right]^{d+1/e} + C_{d,M,m} \log\left(3 + \frac{t}{\epsilon}\right) \cdot \left[\log\left(3 + \frac{1}{t}\right)\right]^{d+1/e} \\
&\quad + C_{d,M,m}\left[\log\left(3 + \frac{t}{\epsilon}\right)\right]^{5d/4+2(s-1)} \\
&\quad + C_{d,M,m}\left(3 + \frac{t}{\epsilon}\right)^{1/2}\left[\log\left(3 + \frac{t}{\epsilon}\right)\right]^{2(s-1)}\left[\log\left(3 + \frac{1}{t}\right)\right]^{5d/4+2(s-1)/e},
\end{aligned}$$

where the second inequality also uses the fact that $A_{\mathbf{l}} \le 1$. Combining this bound with (D.41) and (D.42) yields

$$\begin{aligned}
\log N_{[\,]}(\epsilon, B_{p_0}(M,t) - \{f^*\}, \|\cdot\|_{p_0,2}) &\le C_{b,B,d,M,m} \log\left(3 + \frac{t}{\epsilon}\right) \cdot \left[\log\left(3 + \frac{t}{\epsilon}\right)\right]^{d+1/e} \\
&\quad + C_{b,B,d,M,m} \log\left(3 + \frac{t}{\epsilon}\right) \cdot \left[\log\left(3 + \frac{1}{t}\right)\right]^{d+1/e} + C_{b,B,d,M,m}\left[\log\left(3 + \frac{t}{\epsilon}\right)\right]^{5d/4+2(s-1)} \\
&\quad + C_{b,B,d,M,m}\left(3 + \frac{t}{\epsilon}\right)^{1/2}\left[\log\left(3 + \frac{t}{\epsilon}\right)\right]^{2(s-1)}\left[\log\left(3 + \frac{1}{t}\right)\right]^{5d/4+2(s-1)/e}.
\end{aligned}$$

From here on, we can repeat the same argument and computations as in the case $f^* = 0$. The only difference is that the constants now depend on $m = m(f^*)$. As a result, we obtain

$$\|\hat{f}_{n,V}^{d,s} - f^*\|_{p_0,2} = O_p\big(n^{-1/2}(\log n)^{a(d,s)}\big),$$

where the constants underlying O_p depend on b, B, d, V, K, σ, and $m(f^*)$. □

D.9 Proof of Lemma 25

Our proof of Lemma 25 is based on the following lemma. This lemma is a variant of Theorem 32, which itself is a variant of Theorem 1 in Chapter 2. Theorem 32 offers a connection between functions on $[0,1]^m$ that are entirely monotone and right-continuous, and the cumulative distribution functions of finite measures on $[0,1]^m \setminus \{\mathbf{0}\}$. The proof of Lemma 35 will be provided right after the proof of Lemma 25.

Lemma 35. *Suppose a real-valued function g on $[0,1]^m$ is entirely monotone and right-continuous. Also, assume that g is left-continuous at each point $\mathbf{x} \in [0,1]^m \setminus [0,1)^m$ with respect to all the j^{th} coordinates where $x_j = 1$. Then, there exists a finite (Borel) measure ν on $[0,1)^m \setminus \{\mathbf{0}\}$ such that*

$$g(x_1, \dots, x_m) - g(0, \dots, 0) = \nu\biggl(\prod_{j=1}^{m} [0, x_j] \cap \bigl([0,1)^m \setminus \{\mathbf{0}\}\bigr)\biggr) \quad \textit{for } (x_1, \dots, x_m) \in [0,1]^m. \tag{D.43}$$

Theorem 32. *Suppose a real-valued function g on $[0,1]^m$ is entirely monotone and right-continuous. Then, there exists a unique finite (Borel) measure ν on $[0,1]^m \setminus \{\mathbf{0}\}$ such that*

$$g(x_1, \dots, x_m) - g(0, \dots, 0) = \nu\biggl(\prod_{j=1}^{m} [0, x_j] \setminus \{\mathbf{0}\}\biggr) \quad \textit{for } (x_1, \dots, x_m) \in [0,1]^m.$$

Proof of Lemma 25. First, we assume that $f \in \mathcal{F}_{\mathrm{tc}}^{d,s}$ is of the form (4.1) and show that such f can be represented as (D.1). For each $S \subseteq [d]$ with $0 < |S| \le s$, let g_S be the function on $[0,1]^{|S|}$ defined by

$$g_S(x_j, j \in S) = -c_S + \nu_S\biggl(\prod_{j \in S} [0, x_j] \cap \bigl([0,1)^{|S|} \setminus \{\mathbf{0}\}\bigr)\biggr)$$

for $(x_j, j \in S) \in [0,1]^{|S|}$. It can be readily checked that g_S is right-continuous on $[0,1]^{|S|}$ and left-continuous at each point $(x_j, j \in S) \in [0,1]^{|S|} \setminus [0,1)^{|S|}$ with respect to all the j^{th} coordinates where $x_j = 1$. Moreover, for each S, we have

$$\begin{aligned}
\int_{[0,1)^{|S|} \setminus \{\mathbf{0}\}} \prod_{j \in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S) &= \int_{\prod_{j \in S}[0, x_j) \setminus \{\mathbf{0}\}} \int_{\prod_{j \in S}[t_j, x_j)} 1 \, d(s_j, j \in S) \, d\nu_S(t_j, j \in S) \\
&= \int_{\prod_{j \in S}[0, x_j)} \int_{\prod_{j \in S}[0, s_j] \setminus \{\mathbf{0}\}} 1 \, d\nu_S(t_j, j \in S) \, d(s_j, j \in S) \\
&= c_S \cdot \prod_{j \in S} x_j + \int_{\prod_{j \in S}[0, x_j]} g_S(s_j, j \in S) \, d(s_j, j \in S),
\end{aligned}$$

from which it follows that

$$f(x_1, \dots, x_d) = c_0 + \sum_{0<|S|\le s} c_S \cdot \prod_{j \in S} x_j - \sum_{0<|S|\le s} \int_{[0,1)^{|S|} \setminus \{\mathbf{0}\}} \prod_{j\in S} (x_j - t_j)_+ \, d\nu_S(t_j, j \in S)$$
$$= c_0 - \sum_{0<|S|\le s} \int_{\prod_{j\in S}[0,x_j]} g_S(s_j, j \in S) \, d(s_j, j \in S).$$

Hence, it remains to verify that g_S is entirely monotone for each S. Without loss of generality, we assume that $S = [m]$. Recall that, when $S = [m]$, g_S is entirely monotone if

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & ; \; g_S \\ v_1^{(m)}, \dots, v_{p_m+1}^{(m)} & \end{bmatrix} \ge 0$$

for every $(p_1, \dots, p_m) \in \{0,1\}^m \setminus \{\mathbf{0}\}$. For each $\mathbf{p} = (p_1, \dots, p_m)$, this inequality follows from the nonnegativity of ν_S and the following equation:

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & ; \; g_S \\ v_1^{(m)}, \dots, v_{p_m+1}^{(m)} & \end{bmatrix} = \frac{1}{\prod_{j \in S_{\mathbf{p}}} (v_2^{(j)} - v_1^{(j)})} \cdot \nu_S\left(\left(\prod_{j \in S_{\mathbf{p}}} (v_1^{(j)}, v_2^{(j)}] \times \prod_{j \notin S_{\mathbf{p}}} [0, v_1^{(j)}] \right) \cap [0,1)^m \right),$$

where $S_{\mathbf{p}} = \{j \in [m] : p_j = 1\}$.

Next, we assume that f is of the form (D.1) and show that $f \in \mathcal{F}_{\mathrm{tc}}^{d,s}$. For each $S \subseteq [d]$ with $0 < |S| \le s$, since g_S appearing in (D.1) satisfies the conditions of Lemma 35, there exists a finite measure ν_S on $[0,1)^{|S|} \setminus \{\mathbf{0}\}$ such that

$$g_S(x_j, j \in S) - g_S(0, j \in S) = \nu_S\left(\prod_{j \in S} [0, x_j] \cap \left([0,1)^{|S|} \setminus \{\mathbf{0}\} \right) \right)$$

for every $(x_j, j \in S) \in [0,1]^{|S|}$. If we let $c_S = -g_S(0, j \in S)$, then we can express f as

$$f(x_1, \dots, x_d) = c_0 + \sum_{0<|S|\le s} c_S \prod_{j \in S} x_j - \sum_{0<|S|\le s} \int_{\prod_{j\in S}[0,x_j)} \nu_S\left(\prod_{j \in S} [0, t_j] \setminus \{\mathbf{0}\} \right) d(t_j, j \in S)$$
$$= c_0 + \sum_{0<|S|\le s} c_S \prod_{j \in S} x_j - \sum_{0<|S|\le s} \int_{\prod_{j\in S}[0,x_j)} \int_{\prod_{j\in S}[0,t_j] \setminus \{\mathbf{0}\}} 1 \, d\nu_S(s_j, j \in S) \, d(t_j, j \in S)$$
$$= c_0 + \sum_{0<|S|\le s} c_S \prod_{j \in S} x_j - \sum_{0<|S|\le s} \int_{\prod_{j\in S}[0,x_j) \setminus \{\mathbf{0}\}} \int_{\prod_{j\in S}[s_j,x_j)} 1 \, d(t_j, j \in S) \, d\nu_S(s_j, j \in S)$$
$$= c_0 + \sum_{0<|S|\le s} c_S \prod_{j \in S} x_j - \sum_{0<|S|\le s} \int_{[0,1)^{|S|} \setminus \{\mathbf{0}\}} \prod_{j \in S} (x_j - s_j)_+ \, d\nu_S(s_j, j \in S),$$

which proves that $f \in \mathcal{F}_{\mathrm{tc}}^{d,s}$. □

Proof of Lemma 35. Since g satisfies the conditions of Theorem 32, there exists a finite measure $\bar{\nu}$ on $[0,1]^m \setminus \{\mathbf{0}\}$ such that

$$g(x_1, \dots, x_m) - g(0, \dots, 0) = \bar{\nu}\bigg(\prod_{j=1}^{m}[0, x_j] \setminus \{\mathbf{0}\}\bigg)$$

for every $(x_1, \dots, x_m) \in [0,1]^m$. Let ν be the measure obtained by restricting $\bar{\nu}$ to $[0,1)^m \setminus \{\mathbf{0}\}$. We show that ν is the desired measure, that is, ν satisfies equation (D.43) for every $\mathbf{x} \in [0,1]^m$. For $\mathbf{x} \in [0,1)^m$, (D.43) is clear from the fact that

$$\bar{\nu}\bigg(\prod_{j=1}^{m}[0, x_j] \setminus \{\mathbf{0}\}\bigg) = \bar{\nu}\bigg(\prod_{j=1}^{m}[0, x_j] \cap \big([0,1)^m \setminus \{\mathbf{0}\}\big)\bigg).$$

Hence, we only need to prove (D.43) for $\mathbf{x} \in [0,1]^m \setminus [0,1)^m$. Without loss of generality, let $\mathbf{x} = (1, \dots, 1, x_{l+1}, \dots, x_m)$ where $x_{l+1}, \dots, x_m < 1$. Then, using the left-continuity of g at $\mathbf{x}$, we can show that

$$\begin{aligned} g(x_1, \dots, x_m) - g(0, \dots, 0) &= \lim_{t_1 \uparrow 1} \dots \lim_{t_l \uparrow 1} g(t_1, \dots, t_l, x_{l+1}, \dots, x_m) - g(0, \dots, 0) \\ &= \lim_{t_1 \uparrow 1} \dots \lim_{t_l \uparrow 1} \bar{\nu}\bigg(\prod_{j=1}^{l}[0, t_j] \times \prod_{j=l+1}^{m}[0, x_j] \setminus \{\mathbf{0}\}\bigg) = \bar{\nu}\bigg([0,1)^l \times \prod_{j=l+1}^{m}[0, x_j] \setminus \{\mathbf{0}\}\bigg) \\ &= \bar{\nu}\bigg(\prod_{j=1}^{m}[0, x_j] \cap \big([0,1)^m \setminus \{\mathbf{0}\}\big)\bigg) = \nu\bigg(\prod_{j=1}^{m}[0, x_j] \cap \big([0,1)^m \setminus \{\mathbf{0}\}\big)\bigg). \end{aligned}$$

□

D.10 Proof of Lemma 26

Proof of Lemma 26. We prove that the divided difference of f of order $\mathbf{p}$ is nonpositive on $[0,1]^d$ for every $\mathbf{p} \in \{0,1,2\}^d$ with $\max_j p_j = 2$ by induction on $|\mathbf{p}| := |\{j : p_j \neq 0\}|$. It is clear that the claim is true when $|\mathbf{p}| = d$, since $N^{(\mathbf{p})} = [0,1]^d$ in this case. Suppose that the claim is true when $|\mathbf{p}| = m+1$ for some $m < d$ and fix a nonnegative integer vector $\mathbf{p} \in \{0,1,2\}^d$ with $\max_j p_j = 2$ and $|\mathbf{p}| = m$. Without loss of generality, we assume that $\{j : p_j \neq 0\} = [m]$. Then, for every $0 \le v_1^{(j)} < \dots < v_{p_j+1}^{(j)} \le 1$, $j \in [d]$, we have

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & \\ v_1^{(m)}, \dots, v_{p_m+1}^{(m)} & \\ v_1^{(m+1)} & ; \ f \\ \vdots & \\ v_1^{(d-1)} & \\ v_1^{(d)} & \end{bmatrix} = v_1^{(d)} \cdot \begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & \\ v_1^{(m)}, \dots, v_{p_m+1}^{(m)} & \\ v_1^{(m+1)} & ; \ f \\ \vdots & \\ v_1^{(d-1)} & \\ v_1^{(d)}, 0 & \end{bmatrix} + \begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & \\ v_1^{(m)}, \dots, v_{p_m+1}^{(m)} & \\ v_1^{(m+1)} & ; \ f \\ \vdots & \\ v_1^{(d-1)} & \\ 0 & \end{bmatrix}$$

$$\leq \begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & \\ v_1^{(m)}, \dots, v_{p_m+1}^{(m)} & \\ v_1^{(m+1)} & ; \ f \\ \vdots & \\ v_1^{(d-1)} & \\ 0 & \end{bmatrix},$$

where the inequality is due to the fact that the divided difference of f of order

$$(p_1, \dots, p_m, 0, \dots, 0, 1)$$

is nonpositive because $|(p_1, \dots, p_m, 0, \dots, 0, 1)| = m + 1$. If we repeat this argument with the induction hypothesis that the divided differences of f of order

$$(p_1, \dots, p_m, 0, \dots, 0, 1, 0), \dots, (p_1, \dots, p_m, 1, 0, \dots, 0)$$

are nonpositive, we can show that

$$\begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & \\ v_1^{(m)}, \dots, v_{p_m+1}^{(m)} & ; \ f \\ v_1^{(m+1)} & \\ \vdots & \\ v_1^{(d)} & \end{bmatrix} \leq \begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} & \\ \vdots & \\ v_1^{(m)}, \dots, v_{p_m+1}^{(m)} & ; \ f \\ 0 & \\ \vdots & \\ 0 & \end{bmatrix}.$$

Since the divided difference of f of order $(p_1, \dots, p_m, 0, \dots, 0)$ is nonpositive on

$$N^{(p_1, \dots, p_m, 0, \dots, 0)} = [0, 1]^m \times \{0\} \times \cdots \times \{0\}$$

by the assumption, the right-hand side of this inequality is nonpositive. This proves that the divided difference of f of order $\mathbf{p}$ is nonpositive on $[0, 1]^d$ and hence concludes the proof. □

D.11 Proof of Lemma 27

Proof of Lemma 27. The last part is a direct consequence of the first two. This is because

$$\begin{bmatrix} v_1^{(1)}, v_2^{(1)} & \\ \vdots & ; \ f \\ v_1^{(d)}, v_2^{(d)} & \end{bmatrix} \geq \begin{bmatrix} v_1^{(1)}, w_2^{(1)} & \\ \vdots & ; \ f \\ v_1^{(d)}, w_2^{(d)} & \end{bmatrix} \geq \begin{bmatrix} w_1^{(1)}, w_2^{(1)} & \\ \vdots & ; \ f \\ w_1^{(d)}, w_2^{(d)} & \end{bmatrix}$$

since $v_1^{(j)} \leq w_1^{(j)}$ and $v_2^{(j)} \leq w_2^{(j)}$ for $j \in [d]$.

Suppose we are given $v_1^{(j)} < v_2^{(j)} \leq v_3^{(j)}$ for $j \in [d]$. Since the divided difference of f of order $(1, \ldots, 1, 2)$ is nonpositive, we have the inequality

$$\begin{aligned}
\begin{bmatrix} \begin{matrix} v_1^{(1)}, v_2^{(1)} \\ \vdots \\ v_1^{(d-1)}, v_2^{(d-1)} \\ v_1^{(d)}, v_2^{(d)} \end{matrix} & ; \; f \end{bmatrix} &= -\big(v_3^{(d)} - v_2^{(d)}\big) \cdot \begin{bmatrix} \begin{matrix} v_1^{(1)}, v_2^{(1)} \\ \vdots \\ v_1^{(d-1)}, v_2^{(d-1)} \\ v_1^{(d)}, v_2^{(d)}, v_3^{(d)} \end{matrix} & ; \; f \end{bmatrix} + \begin{bmatrix} \begin{matrix} v_1^{(1)}, v_2^{(1)} \\ \vdots \\ v_1^{(d-1)}, v_2^{(d-1)} \\ v_1^{(d)}, v_3^{(d)} \end{matrix} & ; \; f \end{bmatrix} \\
&\geq \begin{bmatrix} \begin{matrix} v_1^{(1)}, v_2^{(1)} \\ \vdots \\ v_1^{(d-1)}, v_2^{(d-1)} \\ v_1^{(d)}, v_3^{(d)} \end{matrix} & ; \; f \end{bmatrix}.
\end{aligned}$$

Applying the same argument to each coordinate in turn and leveraging the assumption that the divided differences of f of order $(1, \ldots, 1, 2, 1), \ldots, (2, 1, \ldots, 1)$ are nonpositive, we can derive that

$$\begin{bmatrix} \begin{matrix} v_1^{(1)}, v_2^{(1)} \\ \vdots \\ v_1^{(d-1)}, v_2^{(d-1)} \\ v_1^{(d)}, v_2^{(d)} \end{matrix} & ; \; f \end{bmatrix} \geq \begin{bmatrix} \begin{matrix} v_1^{(1)}, v_2^{(1)} \\ \vdots \\ v_1^{(d-1)}, v_2^{(d-1)} \\ v_1^{(d)}, v_3^{(d)} \end{matrix} & ; \; f \end{bmatrix} \geq \cdots \geq \begin{bmatrix} \begin{matrix} v_1^{(1)}, v_3^{(1)} \\ \vdots \\ v_1^{(d-1)}, v_3^{(d-1)} \\ v_1^{(d)}, v_3^{(d)} \end{matrix} & ; \; f \end{bmatrix},$$

as desired. By the same argument, we can show that

$$\begin{bmatrix} \begin{matrix} v_1^{(1)}, v_3^{(1)} \\ \vdots \\ v_1^{(d)}, v_3^{(d)} \end{matrix} & ; \; f \end{bmatrix} \geq \begin{bmatrix} \begin{matrix} v_2^{(1)}, v_3^{(1)} \\ \vdots \\ v_2^{(d)}, v_3^{(d)} \end{matrix} & ; \; f \end{bmatrix}$$

when we are given $v_1^{(j)} \leq v_2^{(j)} < v_3^{(j)}$ for $j \in [d]$. $\square$

D.12 Proof of Lemma 28

Proof of Lemma 28. Without loss of generality, we assume that $\mathbf{p} = (1, \ldots, 1, 0, \ldots, 0)$ and $m = |\{j : p_j \neq 0\}|$. Fix a sufficiently small $\epsilon > 0$ and, for each $j \in [d]$, let $u_1^{(j)} = w_1^{(j)} = v_1^{(j)}$, $u_2^{(j)} = w_2^{(j)} = v_1^{(j)} + \epsilon$ and, if $p_j = 1$, further let $w_3^{(j)} = v_2^{(j)} - \epsilon$ and $u_3^{(j)} = w_4^{(j)} = v_2^{(j)}$. Since the divided difference of f of order $(p_1 + 1, \ldots, p_d + 1) = (2, \ldots, 2, 1, \ldots, 1)$ is nonpositive, we have the inequality

$$
\begin{aligned}
0 \geq & \begin{bmatrix} u_1^{(1)}, u_2^{(1)}, u_3^{(1)} & \\ \vdots & \\ u_1^{(m)}, u_2^{(m)}, u_3^{(m)} & ; \ f \\ u_1^{(m+1)}, u_2^{(m+1)} & \\ \vdots & \\ u_1^{(d)}, u_2^{(d)} & \end{bmatrix} \\
= & \sum_{i_1=1}^{3} \cdots \sum_{i_m=1}^{3} \sum_{i_{m+1}=1}^{2} \cdots \sum_{i_d=1}^{2} \frac{1}{\prod_{l_1 \neq i_1}(u_{i_1}^{(1)} - u_{l_1}^{(1)})} \times \cdots \times \frac{1}{\prod_{l_m \neq i_m}(u_{i_m}^{(m)} - u_{l_m}^{(m)})} \\
& \qquad \cdot \frac{1}{\epsilon^{d-m}} \cdot (-1)^{\sum_{j=m+1}^{d} i_j} \cdot f\left(u_{i_1}^{(1)}, \ldots, u_{i_d}^{(d)}\right) \\
= & \frac{1}{\prod_{j=1}^{m}(u_3^{(j)} - u_1^{(j)})} \cdot \frac{1}{\epsilon^{d-m}} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} (-1)^{\sum_{j=1}^{m} \delta_j} \cdot \frac{1}{\prod_{j=1}^{m}(u_{3-\delta_j}^{(j)} - u_{2-\delta_j}^{(j)})} \\
& \cdot \sum_{i_1=2-\delta_1}^{3-\delta_1} \cdots \sum_{i_m=2-\delta_m}^{3-\delta_m} \sum_{i_{m+1}=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{\sum_{j=1}^{m}(i_j+\delta_j-1)} \cdot (-1)^{\sum_{j=m+1}^{d} i_j} \cdot f\left(u_{i_1}^{(1)}, \ldots, u_{i_d}^{(d)}\right) \\
= & \frac{1}{\prod_{j=1}^{m}(u_3^{(j)} - u_1^{(j)})} \cdot \frac{1}{\epsilon^{d-m}} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} (-1)^{\sum_{j=1}^{m} \delta_j} \cdot \frac{1}{\prod_{j=1}^{m}(u_{3-\delta_j}^{(j)} - u_{2-\delta_j}^{(j)})} \sum_{i_1=2-\delta_1}^{3-\delta_1} \cdots \sum_{i_m=2-\delta_m}^{3-\delta_m} \\
& \sum_{i_{m+1}=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{\sum_{j=1}^{m}(i_j+\delta_j-1)} \cdot (-1)^{\sum_{j=m+1}^{d} i_j} \cdot f\left(u_{i_1}^{(1)}, \ldots, u_{i_m}^{(m)}, w_{i_{m+1}}^{(m+1)}, \ldots, w_{i_d}^{(d)}\right).
\end{aligned} \tag{D.44}
$$

Also, we have

$$
\begin{aligned}
\text{(D.44)} \geq & \frac{1}{\prod_{j=1}^{m}(u_3^{(j)} - u_1^{(j)})} \cdot \frac{1}{\epsilon^{d-m}} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} (-1)^{\sum_{j=1}^{m} \delta_j} \\
& \cdot \frac{1}{\prod_{j=1}^{m-1}(u_{3-\delta_j}^{(j)} - u_{2-\delta_j}^{(j)})} \cdot \frac{1}{w_{4-2\delta_m}^{(m)} - w_{3-2\delta_m}^{(m)}} \sum_{i_1=2-\delta_1}^{3-\delta_1} \cdots \sum_{i_{m-1}=2-\delta_{m-1}}^{3-\delta_{m-1}} \sum_{i_m=3-2\delta_m}^{4-2\delta_m} \sum_{i_{m+1}=1}^{2} \cdots \sum_{i_d=1}^{2} \\
& (-1)^{\sum_{j=1}^{m-1}(i_j+\delta_j-1)} \cdot (-1)^{\sum_{j=m}^{d} i_j} \cdot f\left(u_{i_1}^{(1)}, \ldots, u_{i_{m-1}}^{(m-1)}, w_{i_m}^{(m)}, \ldots, w_{i_d}^{(d)}\right),
\end{aligned} \tag{D.45}
$$

which is because

$$
\begin{aligned}
\text{(D.45)} - \text{(D.44)} = & \frac{w_3^{(m)} - w_2^{(m)}}{u_3^{(m)} - u_1^{(m)}} \cdot \frac{1}{\prod_{j=1}^{m-1}(u_3^{(j)} - u_1^{(j)})} \cdot \frac{1}{w_4^{(m)} - w_2^{(m)}} \cdot \frac{1}{\epsilon^{d-m}} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} (-1)^{\sum_{j=1}^{m} \delta_j} \\
& \cdot \frac{1}{\prod_{j=1}^{m-1}(u_{3-\delta_j}^{(j)} - u_{2-\delta_j}^{(j)})} \cdot \frac{1}{w_{4-\delta_m}^{(m)} - w_{3-\delta_m}^{(m)}} \cdot \sum_{i_1=2-\delta_1}^{3-\delta_1} \cdots \sum_{i_{m-1}=2-\delta_{m-1}}^{3-\delta_{m-1}} \sum_{i_m=3-\delta_m}^{4-\delta_m} \sum_{i_{m+1}=1}^{2} \cdots \sum_{i_d=1}^{2}
\end{aligned}
$$

$$(-1)^{\sum_{j=1}^{m-1}(i_j+\delta_j-1)} \cdot (-1)^{i_m+\delta_m} \cdot (-1)^{\sum_{j=m+1}^{d} i_j} \cdot f\big(u_{i_1}^{(1)}, \ldots, u_{i_{m-1}}^{(m-1)}, w_{i_m}^{(m)}, \ldots, w_{i_d}^{(d)}\big)$$

$$= \frac{w_3^{(m)} - w_2^{(m)}}{u_3^{(m)} - u_1^{(m)}} \cdot \begin{bmatrix} \begin{array}{c} u_1^{(1)}, u_2^{(1)}, u_3^{(1)} \\ \vdots \\ u_1^{(m-1)}, u_2^{(m-1)}, u_3^{(m-1)} \\ w_2^{(m)}, w_3^{(m)}, w_4^{(m)} \\ u_1^{(m+1)}, u_2^{(m+1)} \\ \vdots \\ u_1^{(d)}, u_2^{(d)} \end{array} & ; \ f \end{bmatrix} \leq 0.$$

Since $w_4^{(m)} - w_3^{(m)} = w_2^{(m)} - w_1^{(m)} = \epsilon$, it follows that

$$0 \geq \text{(D.45)} = \frac{1}{\prod_{j=1}^{m}(u_3^{(j)} - u_1^{(j)})} \cdot \frac{1}{\epsilon^{d-m+1}} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} (-1)^{\sum_{j=1}^{m} \delta_j} \cdot \frac{1}{\prod_{j=1}^{m-1}(u_{3-\delta_j}^{(j)} - u_{2-\delta_j}^{(j)})}$$
$$\cdot \sum_{i_1=2-\delta_1}^{3-\delta_1} \cdots \sum_{i_{m-1}=2-\delta_{m-1}}^{3-\delta_{m-1}} \sum_{i_m=3-2\delta_m}^{4-2\delta_m} \sum_{i_{m+1}=1}^{2} \cdots \sum_{i_d=1}^{2}$$
$$(-1)^{\sum_{j=1}^{m-1}(i_j+\delta_j-1)} \cdot (-1)^{\sum_{j=m}^{d} i_j} \cdot f\big(u_{i_1}^{(1)}, \ldots, u_{i_{m-1}}^{(m-1)}, w_{i_m}^{(m)}, \ldots, w_{i_d}^{(d)}\big).$$

If we apply the same argument to the $(m-1)^{\text{th}}, \ldots, 1^{\text{st}}$ coordinates sequentially, we obtain

$$0 \geq \frac{1}{\prod_{j=1}^{m}(u_3^{(j)} - u_1^{(j)})} \cdot \frac{1}{\epsilon^{d}} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} (-1)^{\sum_{j=1}^{m} \delta_j}$$
$$\cdot \sum_{i_1=3-2\delta_1}^{4-2\delta_1} \cdots \sum_{i_m=3-2\delta_m}^{4-2\delta_m} \sum_{i_{m+1}=1}^{2} \cdots \sum_{i_d=1}^{2} f\big(w_{i_1}^{(1)}, \ldots, w_{i_d}^{(d)}\big)$$
$$= \frac{1}{\prod_{j=1}^{m}(v_2^{(j)} - v_1^{(j)})} \cdot \frac{1}{\epsilon^{d}} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} (-1)^{\sum_{j=1}^{m} \delta_j} \triangle f\left(\prod_{j=1}^{m} \big[w_{3-2\delta_j}^{(j)}, w_{4-2\delta_j}^{(j)}\big] \times \prod_{j=m+1}^{d} \big[w_1^{(j)}, w_2^{(j)}\big]\right).$$

If we take the limit $\epsilon \to 0$, this gives that

$$0 \geq \frac{1}{\prod_{j=1}^{m}(v_2^{(j)} - v_1^{(j)})} \sum_{i_1=1}^{2} \cdots \sum_{i_m=1}^{2} (-1)^{\sum_{j=1}^{m} i_j} \cdot D\triangle f\big(v_{i_1}^{(1)}, \ldots, v_{i_m}^{(m)}, v_1^{(m+1)}, \ldots, v_1^{(d)}\big)$$
$$= \begin{bmatrix} \begin{array}{c} v_1^{(1)}, v_2^{(1)} \\ \vdots \\ v_1^{(m)}, v_2^{(m)} \\ v_1^{(m+1)} \\ \vdots \\ v_1^{(d)} \end{array} & ; \ D\triangle f \end{bmatrix}.$$

□

D.13 Proof of Lemma 31

Proof of Lemma 31. The following lemma will be repeatedly used in our proof of Lemma 31. This lemma can be viewed as a generalization of Lemma 27, and it can be proved similarly to Lemma 27.

Lemma 36. *Suppose $f \in \mathcal{F}^d_{tcp}$. Then, for every nonempty subset $S \subseteq [d]$, we have*

$$\begin{bmatrix} v_1^{(j)}, v_2^{(j)}, & j \in S \\ v_1^{(j)}, & j \notin S \end{bmatrix}; \; f \Bigg] \geq \begin{bmatrix} w_1^{(j)}, w_2^{(j)}, & j \in S \\ w_1^{(j)}, & j \notin S \end{bmatrix}; \; f \Bigg]$$

provided that $0 \leq v_1^{(j)} \leq w_1^{(j)} \leq 1$ and $0 \leq v_2^{(j)} \leq w_2^{(j)} \leq 1$ for $j \in S$, and $0 \leq v_1^{(j)} = w_1^{(j)} \leq 1$ for $j \notin S$.

Also, note that since g^* satisfies the interaction restriction condition (4.16) for every subset $S \subseteq [d]$ with $|S| > s$, in the definition (D.22) of $M(g^*)$, we can equivalently take the maximum over all nonempty subsets $S \subseteq [d]$ as follows:

$$M = M(g^*) = \max_{\emptyset \neq S \subseteq [d]} \max \left(\left| \lim_{t \to 0+} \begin{bmatrix} 0, t, & j \in S \\ 0, & j \notin S \end{bmatrix}; \; g^* \Bigg] \right|, \lim_{t \to 1-} \left| \begin{bmatrix} t, 1, & j \in S \\ 0, & j \notin S \end{bmatrix}; \; g^* \Bigg] \right| \right). \tag{D.46}$$

Before proving (D.23) and (D.24), we show that there exists some constant B_d depending on d such that

$$\left| \begin{bmatrix} v_1^{(1)}, \dots, v_{p_1+1}^{(1)} \\ \vdots \\ v_1^{(d)}, \dots, v_{p_d+1}^{(d)} \end{bmatrix}; g^* \Bigg] \right| \leq B_d M \tag{D.47}$$

for every $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1\}^d \setminus \{\mathbf{0}\}$ and $0 \leq v_1^{(j)} < \cdots < v_{p_j+1}^{(j)} \leq 1$ for $j \in [d]$. Since $\boldsymbol{\theta}^* = (g^*(\mathbf{x}^{(\mathbf{i})}), \mathbf{i} \in I_0)$, this result directly implies that

$$\left| \left(\prod_{j=1}^d n_j^{p_j} \right) \cdot (D^{(\mathbf{p})} \theta^*)_{\mathbf{i}} \right| \leq B_d M \tag{D.48}$$

for every $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1\}^d \setminus \{\mathbf{0}\}$ and $\mathbf{i} = (i_1, \dots, i_d) \in I_0$ with $i_j \geq p_j$ for $j \in [d]$.

We prove (D.47) by induction on $|\mathbf{p}| := |\{j : p_j \neq 0\}|$. When $|\mathbf{p}| = d$ ($\mathbf{p} = (1, \dots, 1)$), it directly follows from Lemma 36 and (D.46) that the inequality (D.47) holds with $B_d = 1 =: B_{d,d}$. Suppose that the inequality (D.47) holds with $B_d = B_{d,m+1}$ for every $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1\}^d$ with $|\mathbf{p}| = m + 1$, and fix $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1\}^d$ with $|\mathbf{p}| = m$. Without loss of generality, let $\mathbf{p} = (1, \dots, 1, 0, \dots, 0)$. Observe that

$$
\begin{aligned}
\left| \begin{bmatrix} v_1^{(1)}, v_2^{(1)} & \\ \vdots & \\ v_1^{(m)}, v_2^{(m)} & ; \; g^* \\ v_1^{(m+1)} & \\ \vdots & \\ v_1^{(d)} & \end{bmatrix} \right| &\le \left| \begin{bmatrix} v_1^{(1)}, v_2^{(1)} & \\ \vdots & \\ v_1^{(m)}, v_2^{(m)} & \\ 0 & ; g^* \\ v_1^{(m+2)} & \\ \vdots & \\ v_1^{(d)} & \end{bmatrix} \right| + v_1^{(m+1)} \cdot \left| \begin{bmatrix} v_1^{(1)}, v_2^{(1)} & \\ \vdots & \\ v_1^{(m)}, v_2^{(m)} & \\ 0, v_1^{(m+1)} & ; g^* \\ v_1^{(m+2)} & \\ \vdots & \\ v_1^{(d)} & \end{bmatrix} \right| \\
&\le \left| \begin{bmatrix} v_1^{(1)}, v_2^{(1)} & \\ \vdots & \\ v_1^{(m)}, v_2^{(m)} & \\ 0 & ; g^* \\ v_1^{(m+2)} & \\ \vdots & \\ v_1^{(d)} & \end{bmatrix} \right| + B_{d,m+1} M.
\end{aligned}
$$

Repeating this argument with the $(m+2)^{\text{th}}, \dots, d^{\text{th}}$ coordinates, we obtain

$$
\begin{aligned}
\left| \begin{bmatrix} v_1^{(1)}, v_2^{(1)} & \\ \vdots & \\ v_1^{(m)}, v_2^{(m)} & ; \; g^* \\ v_1^{(m+1)} & \\ \vdots & \\ v_1^{(d)} & \end{bmatrix} \right| &\le \left| \begin{bmatrix} v_1^{(1)}, v_2^{(1)} & \\ \vdots & \\ v_1^{(m)}, v_2^{(m)} & ; \; g^* \\ 0 & \\ \vdots & \\ 0 & \end{bmatrix} \right| + (d-m) \cdot B_{d,m+1} M \\
&\le M + (d-m) \cdot B_{d,m+1} M = \big(1 + (d-m) \cdot B_{d,m+1}\big) M,
\end{aligned}
$$

where the second inequality is again from Lemma 36 and (D.46). This proves that the inequality (D.47) holds with $B_d = 1 + (d-m) \cdot B_{d,m+1} =: B_{d,m}$ for every $\mathbf{p} = (p_1, \dots, p_d) \in \{0,1\}^d$ with $|\mathbf{p}| = m$. Consequently, the inequality (D.47) holds for every $\mathbf{p} \in \{0,1\}^d \setminus \{\mathbf{0}\}$, if $B_d = \max_{m \in [d]} B_{d,m}$.

Now, we are ready to prove (D.23) and (D.24). We also prove them by induction on $|\mathbf{p}|$. Let us first consider the case $|\mathbf{p}| = d$ ($\mathbf{p} = (1, \dots, 1)$). Suppose, for contradiction, that there exist $\mathbf{r} = (r_1, \dots, r_d) \in \prod_{j=1}^d \{1, \dots, R_j\}$ and $\mathbf{i} = (i_1, \dots, i_d) \in L_{\mathbf{r}}$ such that

$$
(D^{(\mathbf{p})}\theta)_{\mathbf{i}} < \frac{C_d}{\prod_{j=1}^d n_j} \cdot \Big[-M - t \Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^d r_j} \Big],
$$

where $C_d = \max((2\sqrt{3})^d, B_d) =: C^L_{d,d}$. Then, by the definition (D.17) of discrete differences and Lemma 36, for every $\mathbf{l} = (l_1, \dots, l_d) \in I_0$ with $l_j \geq i_j$ for $j \in [d]$, we have

$$\begin{aligned}(D^{(\mathbf{p})}\theta)_{\mathbf{l}} \leq (D^{(\mathbf{p})}\theta)_{\mathbf{i}} &< \frac{C_d}{\prod_{j=1}^d n_j} \cdot \Big[-M - t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^d r_j}\Big] \\ &\leq (D^{(\mathbf{p})}\theta^*)_{\mathbf{l}} - \frac{C_d}{\prod_{j=1}^d n_j} \cdot t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^d r_j}.\end{aligned}$$

Here, for the last inequality, we use (D.48) and the fact that $C_d \geq B_d$. It thus follows that if $i_1 - 1 \leq l_1' < l_1, \dots, i_d - 1 \leq l_d' < l_d$, then

$$\begin{aligned}&\sum_{\boldsymbol{\delta} \in \{0,1\}^d} (\theta - \theta^*)^2_{(1-\delta_1)l_1 + \delta_1 l_1', \dots, (1-\delta_d)l_d + \delta_d l_d'} \\ &\quad \geq \frac{1}{2^d} \cdot \Bigg(\sum_{\boldsymbol{\delta} \in \{0,1\}^d} (-1)^{\delta_1 + \dots + \delta_d} \cdot (\theta - \theta^*)_{(1-\delta_1)l_1 + \delta_1 l_1', \dots, (1-\delta_d)l_d + \delta_d l_d'} \Bigg)^2 \\ &\quad = \frac{1}{2^d} \cdot \Bigg(\sum_{k_1 = l_1'+1}^{l_1} \cdots \sum_{k_d = l_d'+1}^{l_d} \big(D^{(\mathbf{p})}\theta - D^{(\mathbf{p})}\theta^*\big)_{k_1, \dots, k_d} \Bigg)^2 \\ &\quad > (l_1 - l_1')^2 \times \cdots \times (l_d - l_d')^2 \cdot \frac{C_d^2/2^d}{\prod_{j=1}^d n_j^2} \cdot t^2 \Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j}.\end{aligned}$$

This implies that

$$\begin{aligned}t^2 \geq \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 &\geq \sum_{(i_1+n_1-2)/2 < l_1 \leq n_1 - 1} \cdots \sum_{(i_d+n_d-2)/2 < l_d \leq n_d - 1} \\ &\qquad \sum_{\boldsymbol{\delta} \in \{0,1\}^d} (\theta - \theta^*)^2_{(1-\delta_1)l_1 + \delta_1(i_1+n_1-2-l_1), \dots, (1-\delta_d)l_d + \delta_d(i_d+n_d-2-l_d)} \\ &> \sum_{(i_1+n_1-2)/2 < l_1 \leq n_1 - 1} \cdots \sum_{(i_d+n_d-2)/2 < l_d \leq n_d - 1} \\ &\qquad (2l_1 - i_1 - n_1 + 2)^2 \times \cdots \times (2l_d - i_d - n_d + 2)^2 \cdot \frac{C_d^2/2^d}{\prod_{j=1}^d n_j^2} \cdot t^2 \Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j} \\ &= \prod_{j=1}^d \Big[\frac{1}{6}(n_j - i_j)(n_j - i_j + 1)(n_j - i_j + 2)\Big] \cdot \frac{C_d^2/2^d}{\prod_{j=1}^d n_j^2} \cdot t^2 \Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j} \\ &> \prod_{j=1}^d (n_j - i_j)^3 \cdot \frac{C_d^2/12^d}{\prod_{j=1}^d n_j^2} \cdot t^2 \Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j} \\ &> \prod_{j=1}^d \Big(\frac{n_j}{2^{r_j}}\Big)^3 \cdot \frac{C_d^2/12^d}{\prod_{j=1}^d n_j^2} \cdot t^2 \Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j} = \frac{C_d^2}{12^d} \cdot t^2 \geq t^2,\end{aligned}$$

which leads to contradiction. The second-to-last inequality is from the fact that, for each j,

$$n_j - i_j \geq n_j - \frac{n_j}{2^{r_j}} + 1 > \frac{n_j}{2^{r_j}},$$

since $r_j \geq 1$.

Next, suppose, for contradiction, that there exist $\mathbf{r} = (r_1, \dots, r_d) \in \prod_{j=1}^d \{0, 1, \dots, R_j\}$ and $\mathbf{i} = (i_1, \dots, i_d) \in L_{\mathbf{r}}$ with $i_j \geq 1 = p_j$ for $j \in [d]$ such that

$$(D^{(\mathbf{p})}\theta)_{\mathbf{i}} > \frac{C_d}{\prod_{j=1}^d n_j} \cdot \left[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^d r_j}\right],$$

where $C_d = \max((8\sqrt{2})^d, B_d) =: C_{d,d}^U$. Then, for every $\mathbf{l} = (l_1, \dots, l_d) \in I_0$ with $l_j \leq i_j$,

$$\begin{aligned}(D^{(\mathbf{p})}\theta)_{\mathbf{l}} \geq (D^{(\mathbf{p})}\theta)_{\mathbf{i}} &> \frac{C_d}{\prod_{j=1}^d n_j} \cdot \left[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^d r_j}\right] \\ &\geq (D^{(\mathbf{p})}\theta^*)_{\mathbf{l}} + \frac{C_d}{\prod_{j=1}^d n_j} \cdot t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^d r_j}.\end{aligned}$$

By the same argument as above, we can show that

$$\begin{aligned}&\sum_{\boldsymbol{\delta} \in \{0,1\}^d} (\theta - \theta^*)^2_{(1-\delta_1)l_1 + \delta_1 l_1', \dots, (1-\delta_d)l_d + \delta_d l_d'} \\ &\qquad > (l_1 - l_1')^2 \times \cdots \times (l_d - l_d')^2 \cdot \frac{C_d^2/2^d}{\prod_{j=1}^d n_j^2} \cdot t^2\Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j}\end{aligned}$$

provided that $0 \leq l_1' < l_1 \leq i_1, \dots, 0 \leq l_d' < l_d \leq i_d$. It thus follows that

$$\begin{aligned}t^2 \geq \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 &\geq \sum_{i_1/2 < l_1 \leq i_1} \cdots \sum_{i_d/2 < l_d \leq i_d} \sum_{\boldsymbol{\delta} \in \{0,1\}^d} (\theta - \theta^*)^2_{(1-\delta_1)l_1 + \delta_1(i_1 - l_1), \dots, (1-\delta_d)l_d + \delta_d(i_d - l_d)} \\ &> \sum_{i_1/2 < l_1 \leq i_1} \cdots \sum_{i_d/2 < l_d \leq i_d} (2l_1 - i_1)^2 \times \cdots \times (2l_d - i_d)^2 \cdot \frac{C_d^2/2^d}{\prod_{j=1}^d n_j^2} \cdot t^2\Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j} \\ &= \prod_{j=1}^d \left[\frac{1}{6} i_j(i_j+1)(i_j+2)\right] \cdot \frac{C_d^2/2^d}{\prod_{j=1}^d n_j^2} \cdot t^2\Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j} \\ &\geq \prod_{j=1}^d (i_j+1)^3 \cdot \frac{C_d^2/16^d}{\prod_{j=1}^d n_j^2} \cdot t^2\Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j} \\ &> \prod_{j=1}^d \Big(\frac{n_j}{2^{r_j+1}}\Big)^3 \cdot \frac{C_d^2/16^d}{\prod_{j=1}^d n_j^2} \cdot t^2\Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^d 2r_j} = \frac{C_d^2}{128^d} \cdot t^2 \geq t^2,\end{aligned}$$

which leads to contradiction. Here, for the fourth inequality, we use

$$\frac{4}{3} i_j(i_j+1)(i_j+2) \geq (i_j+1)^3,$$

which holds since $i_j \geq 1$.

Now, we assume that (D.23) and (D.24) hold with $C_d = C^L_{d,m+1}$ and $C_d = C^U_{d,m+1}$, respectively, for every $\mathbf{p} \in \{0,1\}^d$ with $|\mathbf{p}| = m+1$, and prove that they also hold with some $C_d = C^L_{d,m}$ and $C_d = C^U_{d,m}$ for every $\mathbf{p} \in \{0,1\}^d$ with $|\mathbf{p}| = m$. For notational convenience, here, we only consider the case where $\mathbf{p} = (1,\dots,1,0,\dots,0) := \mathbf{1}_m$, but the same argument applies to other $\mathbf{p} \in \{0,1\}^d$ with $|\mathbf{p}| = m$.

Suppose, for contradiction, that there exist $\mathbf{r} = (r_1,\dots,r_d) \in \prod_{j=1}^d \{1,\dots,R_j\}$ and $\mathbf{i} = (i_1,\dots,i_d) \in L_{\mathbf{r}}$ such that

$$(D^{(\mathbf{1}_m)}\theta)_{\mathbf{i}} < \frac{C_d}{\prod_{j=1}^m n_j} \cdot \Big[-M - t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^m r_j} \Big],$$

where

$$C_d = \frac{3}{2} \cdot (d-m) C^U_{d,m+1} + \max((2\sqrt{3})^m, B_d) =: C^L_{d,m}.$$

Then, by the induction hypothesis, for $\mathbf{l} = (l_1,\dots,l_d) \in I_0$ with $l_1 \geq i_1,\dots,l_m \geq i_m$, and $i_{m+1} \leq l_{m+1} \leq b^{(m+1)}_{r_{m+1}-1},\dots,i_d \leq l_d \leq b^{(d)}_{r_d-1}$, we have

$$\begin{aligned}
(D^{(\mathbf{1}_m)}\theta)_{l_1,\dots,l_d} &= (D^{(\mathbf{1}_m)}\theta)_{l_1,\dots,l_m,i_{m+1},l_{m+2},\dots,l_d} + \sum_{k_{m+1}=i_{m+1}+1}^{l_{m+1}} (D^{(\mathbf{1}_{m+1})}\theta)_{l_1,\dots,l_m,k_{m+1},l_{m+2},\dots,l_d} \\
&\leq (D^{(\mathbf{1}_m)}\theta)_{l_1,\dots,l_m,i_{m+1},l_{m+2},\dots,l_d} + (l_{m+1}-i_{m+1}) \cdot \frac{C^U_{d,m+1}}{\prod_{j=1}^{m+1} n_j} \cdot \Big[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^{m+1} r_j} \Big] \\
&\leq (D^{(\mathbf{1}_m)}\theta)_{l_1,\dots,l_m,i_{m+1},l_{m+2},\dots,l_d} + \frac{3}{2} \cdot \frac{C^U_{d,m+1}}{\prod_{j=1}^m n_j} \cdot \Big[\frac{M}{2^{r_{m+1}}} + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^m r_j} \Big],
\end{aligned}$$

where $\mathbf{1}_{m+1} = (1,\dots,1,0,\dots,0)$ with $|\mathbf{1}_{m+1}| = m+1$. Here, the second inequality is because

$$l_{m+1} - i_{m+1} \leq b^{(m+1)}_{r_{m+1}-1} - i_{m+1} < \frac{n_{m+1}}{2^{r_{m+1}-1}} - 1 - \Big(\frac{n_{m+1}}{2^{r_{m+1}+1}} - 1\Big) = \frac{3}{2} \cdot \frac{n_{m+1}}{2^{r_{m+1}}}.$$

Repeating this argument with the $(m+2)^{\text{th}},\dots,d^{\text{th}}$ coordinates in turn, we can derive that

$$\begin{aligned}
(D^{(\mathbf{1}_m)}\theta)_{l_1,\dots,l_d} &\leq (D^{(\mathbf{1}_m)}\theta)_{l_1,\dots,l_m,i_{m+1},\dots,i_d} \\
&\quad + \frac{3}{2} \cdot \frac{C^U_{d,m+1}}{\prod_{j=1}^m n_j} \cdot \Big[\Big(\sum_{j=m+1}^d \frac{1}{2^{r_j}} \Big) \cdot M + (d-m) \cdot t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^m r_j} \Big] \\
&\leq (D^{(\mathbf{1}_m)}\theta)_{l_1,\dots,l_m,i_{m+1},\dots,i_d} + \frac{3}{2} \cdot (d-m) \cdot \frac{C^U_{d,m+1}}{\prod_{j=1}^m n_j} \cdot \Big[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^m r_j} \Big]
\end{aligned}$$

$$
\begin{aligned}
&\leq (D^{(\mathbf{1}_m)}\theta)_{i_1,\dots,i_m,i_{m+1},\dots,i_d} + \frac{3}{2}\cdot(d-m)\cdot\frac{C^U_{d,m+1}}{\prod_{j=1}^m n_j}\cdot\left[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2}\cdot 2^{\sum_{j=1}^m r_j}\right] \\
&< -\Big(C_d - \frac{3}{2}\cdot(d-m)C^U_{d,m+1}\Big)\cdot\frac{1}{\prod_{j=1}^m n_j}\cdot\left[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2}\cdot 2^{\sum_{j=1}^m r_j}\right] \\
&\leq (D^{(\mathbf{1}_m)}\theta^*)_{l_1,\dots,l_d} - \Big(C_d - \frac{3}{2}\cdot(d-m)C^U_{d,m+1}\Big)\cdot\frac{1}{\prod_{j=1}^m n_j}\cdot t\Big(\frac{2^{r_+}}{n}\Big)^{1/2}\cdot 2^{\sum_{j=1}^m r_j},
\end{aligned}
$$

where the third inequality follows from Lemma 36 and the definition (D.17) of discrete differences, and the last inequality is due to (D.48) and the fact that

$$\widetilde{C}_d := C_d - \frac{3}{2}\cdot(d-m)C^U_{d,m+1} \geq B_d.$$

It thus follows that if $i_1 - 1 \leq l_1' < l_1, \dots, i_m - 1 \leq l_m' < l_m$, and if $i_{m+1} \leq l_{m+1} \leq b^{(m+1)}_{r_{m+1}-1}, \dots, i_d \leq l_d \leq b^{(d)}_{r_d-1}$, then

$$
\begin{aligned}
&\sum_{\boldsymbol{\delta}\in\{0,1\}^m} (\theta-\theta^*)^2_{(1-\delta_1)l_1+\delta_1 l_1',\dots,(1-\delta_m)l_m+\delta_m l_m', l_{m+1},\dots,l_d} \\
&\quad\geq \frac{1}{2^m}\cdot\left(\sum_{\boldsymbol{\delta}\in\{0,1\}^m}(-1)^{\delta_1+\dots+\delta_m}\cdot(\theta-\theta^*)_{(1-\delta_1)l_1+\delta_1 l_1',\dots,(1-\delta_m)l_m+\delta_m l_m', l_{m+1},\dots,l_d}\right)^2 \\
&\quad= \frac{1}{2^m}\cdot\left(\sum_{k_1=l_1'+1}^{l_1}\cdots\sum_{k_m=l_m'+1}^{l_m}\big(D^{(\mathbf{1}_m)}\theta - D^{(\mathbf{1}_m)}\theta^*\big)_{k_1,\dots,k_m,l_{m+1},\dots,l_d}\right)^2 \\
&\quad> (l_1-l_1')^2\times\cdots\times(l_m-l_m')^2\cdot\frac{\widetilde{C}_d^2/2^m}{\prod_{j=1}^m n_j^2}\cdot t^2\Big(\frac{2^{r_+}}{n}\Big)\cdot 2^{\sum_{j=1}^m 2r_j}.
\end{aligned}
$$

This implies that

$$
\begin{aligned}
t^2 &\geq \|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \\
&\geq \sum_{(i_1+n_1-2)/2<l_1\leq n_1-1}\cdots\sum_{(i_m+n_m-2)/2<l_m\leq n_m-1}\ \sum_{l_{m+1}=i_{m+1}}^{b^{(m+1)}_{r_{m+1}-1}}\cdots\sum_{l_d=i_d}^{b^{(d)}_{r_d-1}} \\
&\qquad\sum_{\boldsymbol{\delta}\in\{0,1\}^m}(\theta-\theta^*)^2_{(1-\delta_1)l_1+\delta_1(i_1+n_1-2-l_1),\dots,(1-\delta_m)l_m+\delta_m(i_m+n_m-2-l_m),l_{m+1},\dots,l_d} \\
&> \sum_{(i_1+n_1-2)/2<l_1\leq n_1-1}\cdots\sum_{(i_m+n_m-2)/2<l_m\leq n_m-1}\ \sum_{l_{m+1}=i_{m+1}}^{b^{(m+1)}_{r_{m+1}-1}}\cdots\sum_{l_d=i_d}^{b^{(d)}_{r_d-1}} \\
&\qquad(2l_1-i_1-n_1+2)^2\times\cdots\times(2l_m-i_m-n_m+2)^2\cdot\frac{\widetilde{C}_d^2/2^m}{\prod_{j=1}^m n_j^2}\cdot t^2\Big(\frac{2^{r_+}}{n}\Big)\cdot 2^{\sum_{j=1}^m 2r_j}
\end{aligned}
$$

$$
\begin{aligned}
&= \prod_{j=1}^{m} \Big[\frac{1}{6}(n_j - i_j)(n_j - i_j + 1)(n_j - i_j + 2)\Big] \cdot \prod_{j=m+1}^{d} \big(b^{(j)}_{r_j - 1} - i_j + 1\big) \\
&\qquad \cdot \frac{\widetilde{C}_d^2 / 2^m}{\prod_{j=1}^m n_j^2} \cdot t^2 \Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^m 2r_j} \\
&> \prod_{j=1}^{m} (n_j - i_j)^3 \cdot \prod_{j=m+1}^{d} \big(b^{(j)}_{r_j - 1} - i_j + 1\big) \cdot \frac{\widetilde{C}_d^2 / 12^m}{\prod_{j=1}^m n_j^2} \cdot t^2 \Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^m 2r_j} \\
&> \prod_{j=1}^{m} \Big(\frac{n_j}{2^{r_j}}\Big)^3 \cdot \prod_{j=m+1}^{d} \frac{n_j}{2^{r_j}} \cdot \frac{\widetilde{C}_d^2 / 12^m}{\prod_{j=1}^m n_j^2} \cdot t^2 \Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^m 2r_j} = \frac{\widetilde{C}_d^2}{12^m} \cdot t^2 \geq t^2,
\end{aligned}
$$

which leads to contradiction. Here, the second-to-last inequality is because

$$
b^{(j)}_{r_j - 1} - i_j + 1 > \frac{n_j}{2^{r_j - 1}} - 2 - \Big(\frac{n_j}{2^{r_j}} - 1\Big) + 1 = \frac{n_j}{2^{r_j}} \quad \text{for each } j = m+1, \ldots, d.
$$

Next, suppose, for contradiction, that there exist $\mathbf{r} = (r_1, \ldots, r_d) \in \prod_{j=1}^d \{0, 1, \ldots, R_j\}$ and $\mathbf{i} = (i_1, \ldots, i_d) \in L_{\mathbf{r}}$ with $i_j \geq p_j$ for $j \in [d]$ such that

$$
(D^{(\mathbf{1}_m)}\theta)_{\mathbf{i}} > \frac{C_d}{\prod_{j=1}^m n_j} \cdot \Big[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^m r_j}\Big],
$$

where

$$
C_d = (d-m) C^U_{d,m+1} + \max((4\sqrt{2})^m \cdot 2^d, B_d) =: C^U_{d,m}.
$$

It then follows that, for every $\mathbf{l} = (l_1, \ldots, l_d) \in I_0$ with $l_1 \leq i_1, \ldots, l_m \leq i_m$, and $a^{(m+1)}_{r_{m+1}+1} \leq l_{m+1} \leq i_{m+1}, \ldots, a^{(d)}_{r_d + 1} \leq l_d \leq i_d$ (if $r_j = R_j$, let $a^{(j)}_{R_j + 1} = 0$),

$$
\begin{aligned}
&(D^{(\mathbf{1}_m)}\theta)_{l_1, \ldots, l_d} \geq (D^{(\mathbf{1}_m)}\theta)_{i_1, \ldots, i_m, l_{m+1}, \ldots, l_d} \\
&\quad = (D^{(\mathbf{1}_m)}\theta)_{i_1, \ldots, i_m, i_{m+1}, l_{m+2}, \ldots, l_d} - \sum_{k_{m+1} = l_{m+1}+1}^{i_{m+1}} (D^{(\mathbf{1}_{m+1})}\theta)_{i_1, \ldots, i_m, k_{m+1}, l_{m+2}, \ldots, l_d} \\
&\quad \geq (D^{(\mathbf{1}_m)}\theta)_{i_1, \ldots, i_m, i_{m+1}, l_{m+2}, \ldots, l_d} - (i_{m+1} - l_{m+1}) \cdot \frac{C^U_{d,m+1}}{\prod_{j=1}^{m+1} n_j} \cdot \Big[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^{m+1} r_j}\Big] \\
&\quad \geq (D^{(\mathbf{1}_m)}\theta)_{i_1, \ldots, i_m, i_{m+1}, l_{m+2}, \ldots, l_d} - \frac{C^U_{d,m+1}}{\prod_{j=1}^{m} n_j} \cdot \Big[\frac{M}{2^{r_{m+1}}} + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^{m} r_j}\Big],
\end{aligned}
$$

where the last inequality is from the fact that

$$
i_{m+1} - l_{m+1} \leq i_{m+1} - a^{(m+1)}_{r_{m+1}+1} < \frac{n_{m+1}}{2^{r_{m+1}}} - 1 - \Big(\frac{n_{m+1}}{2^{r_{m+1}+2}} - 1\Big) = \frac{3}{4} \cdot \frac{n_{m+1}}{2^{r_{m+1}}}
$$

when $r_{m+1} \neq R_{m+1}$, and

$$
i_{m+1} - l_{m+1} \leq i_{m+1} \leq \frac{n_{m+1}}{2^{r_{m+1}}} - 1 < \frac{n_{m+1}}{2^{r_{m+1}}}
$$

when $r_{m+1} = R_{m+1}$. If we apply the same argument to the $(m+2)^{\text{th}}, \dots, d^{\text{th}}$ coordinates in turn, we obtain

$$\begin{aligned}
(D^{(\mathbf{1}_m)}\theta)_{l_1,\dots,l_d} &\geq (D^{(\mathbf{1}_m)}\theta)_{i_1,\dots,i_d} - (d-m) \cdot \frac{C^U_{d,m+1}}{\prod_{j=1}^m n_j} \cdot \left[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^m r_j} \right] \\
&> \big(C_d - (d-m)C^U_{d,m+1}\big) \cdot \frac{1}{\prod_{j=1}^m n_j} \cdot \left[M + t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^m r_j} \right] \\
&\geq (D^{(\mathbf{1}_m)}\theta^*)_{l_1,\dots,l_d} + \big(C_d - (d-m)C^U_{d,m+1}\big) \cdot \frac{1}{\prod_{j=1}^m n_j} \cdot t\Big(\frac{2^{r_+}}{n}\Big)^{1/2} \cdot 2^{\sum_{j=1}^m r_j}.
\end{aligned}$$

Here, the last inequality is attributed to (D.48) and the fact that

$$\widetilde{C}_d := C_d - (d-m)C^U_{d,m+1} \geq B_d.$$

Hence, by Cauchy inequality, if $0 \leq l'_1 < l_1 \leq i_1, \dots, 0 \leq l'_m < l_m \leq i_m$, and if $a^{(m+1)}_{r_{m+1}+1} \leq l_{m+1} \leq i_{m+1}, \dots, a^{(d)}_{r_d+1} \leq l_d \leq i_d$, then

$$\begin{aligned}
&\sum_{\boldsymbol{\delta} \in \{0,1\}^m} (\theta - \theta^*)^2_{(1-\delta_1)l_1 + \delta_1 l'_1, \dots, (1-\delta_m)l_m + \delta_m l'_m, l_{m+1}, \dots, l_d} \\
&\qquad > (l_1 - l'_1)^2 \times \dots \times (l_m - l'_m)^2 \cdot \frac{\widetilde{C}_d^2/2^m}{\prod_{j=1}^m n_j^2} \cdot t^2\Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^m 2r_j},
\end{aligned}$$

as above. As a result,

$$\begin{aligned}
t^2 \geq \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 &\geq \sum_{i_1/2 < l_1 \leq i_1} \cdots \sum_{i_m/2 < l_m \leq i_m} \sum_{l_{m+1} = a^{(m+1)}_{r_{m+1}+1}}^{i_{m+1}} \cdots \sum_{l_d = a^{(d)}_{r_d+1}}^{i_d} \\
&\qquad \sum_{\boldsymbol{\delta} \in \{0,1\}^m} (\theta - \theta^*)^2_{(1-\delta_1)l_1 + \delta_1(i_1 - l_1), \dots, (1-\delta_m)l_m + \delta_m(i_m - l_m), l_{m+1}, \dots, l_d} \\
&> \sum_{i_1/2 < l_1 \leq i_1} \cdots \sum_{i_m/2 < l_m \leq i_m} \sum_{l_{m+1} = a^{(m+1)}_{r_{m+1}+1}}^{i_{m+1}} \cdots \sum_{l_d = a^{(d)}_{r_d+1}}^{i_d} \\
&\qquad (2l_1 - i_1)^2 \times \dots \times (2l_m - i_m)^2 \cdot \frac{\widetilde{C}_d^2/2^m}{\prod_{j=1}^m n_j^2} \cdot t^2\Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^m 2r_j} \\
&= \prod_{j=1}^m \Big[\frac{1}{6} i_j(i_j+1)(i_j+2)\Big] \cdot \prod_{j=m+1}^d \big(i_j - a^{(j)}_{r_j+1} + 1\big) \cdot \frac{\widetilde{C}_d^2/2^m}{\prod_{j=1}^m n_j^2} \cdot t^2\Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^m 2r_j} \\
&\geq \prod_{j=1}^m (i_j+1)^3 \cdot \prod_{j=m+1}^d \big(i_j - a^{(j)}_{r_j+1} + 1\big) \cdot \frac{\widetilde{C}_d^2/16^m}{\prod_{j=1}^m n_j^2} \cdot t^2\Big(\frac{2^{r_+}}{n}\Big) \cdot 2^{\sum_{j=1}^m 2r_j}
\end{aligned}$$

$$> \prod_{j=1}^{m} \left(\frac{n_j}{2^{r_j+1}}\right)^3 \cdot \prod_{j=m+1}^{d} \frac{n_j}{2^{r_j+2}} \cdot \frac{\widetilde{C}_d^2/16^m}{\prod_{j=1}^m n_j^2} \cdot t^2\left(\frac{2^{r_+}}{n}\right) \cdot 2^{\sum_{j=1}^m 2r_j} = \frac{\widetilde{C}_d^2}{32^m \cdot 4^d} \cdot t^2 \geq t^2,$$

which leads to contradiction. Here, for the last inequality, we use

$$i_j - a^{(j)}_{r_j+1} + 1 > \frac{n_j}{2^{r_j+1}} - 1 - \frac{n_j}{2^{r_j+2}} + 1 = \frac{n_j}{2^{r_j+2}}$$

for each $j = m+1, \dots, d$.

From these results, we can conclude that (D.23) and (D.24) hold with $C_d = C^L_{d,m}$ and $C_d = C^U_{d,m}$, respectively, for every $\mathbf{p} \in \{0,1\}^d$ with $|\mathbf{p}| = m$. Thus, by induction, (D.23) and (D.24) hold for every $\mathbf{p} \in \{0,1\}^d \setminus \{\mathbf{0}\}$, if we define $C_d = \max(\max_{m\in[d]} C^L_{d,m}, \max_{m\in[d]} C^U_{d,m})$. □

D.14 Proof of Lemma 33

Proof of Lemma 33. The proof proceeds similarly to that of Lemma 31. We prove (D.38) and (D.39) by induction on $|S|$, and Lemma 36 will be repeatedly used in the proof.

We first consider the case where $|S| = d$ ($S = [d]$). Suppose, for contradiction, that there exist $\mathbf{r} = (r_1, \dots, r_d) \in \mathbb{N}^d$ and $1/2^{r_j+1} \leq s_j < t_j \leq 1/2^{r_j}$ for $j \in [d]$ such that

$$\begin{bmatrix} s_1, t_1 & & \\ \vdots & ; & f \\ s_d, t_d & & \end{bmatrix} < -C_d t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^d r_j},$$

where $C_d = (2\sqrt{3})^d =: C^L_{d,d}$. By Lemma 36, we have

$$\begin{bmatrix} v_1, w_1 & & \\ \vdots & ; & f \\ v_d, w_d & & \end{bmatrix} \leq \begin{bmatrix} s_1, t_1 & & \\ \vdots & ; & f \\ s_d, t_d & & \end{bmatrix} < -C_d t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^d r_j}$$

for every $t_j \leq v_j < w_j$ for $j \in [d]$. Applying Cauchy inequality, we thus obtain

$$\begin{aligned} &\sum_{\boldsymbol{\delta}\in\{0,1\}^d} \Big(f\big((1-\delta_1)w_1 + \delta_1 v_1, \dots, (1-\delta_d)w_d + \delta_d v_d\big)\Big)^2 \\ &\geq \frac{1}{2^d}\Bigg(\sum_{\boldsymbol{\delta}\in\{0,1\}^d} (-1)^{\delta_1+\cdots+\delta_d} \cdot f\big((1-\delta_1)w_1 + \delta_1 v_1, \dots, (1-\delta_d)w_d + \delta_d v_d\big)\Bigg)^2 \\ &= \frac{1}{2^d}\prod_{j=1}^d (w_j - v_j)^2 \cdot \begin{bmatrix} v_1, w_1 & & \\ \vdots & ; & f \\ v_d, w_d & & \end{bmatrix}^2 > \prod_{j=1}^d (w_j - v_j)^2 \cdot \frac{C_d^2}{2^d} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^d r_j}. \end{aligned}$$

Since $t_j \leq x_j < t_j + 1 - x_j$ whenever $x_j \in [t_j, (t_j+1)/2)$, it follows that

$$\begin{aligned}
t^2 &\geq \|f\|_2^2 \\
&\geq \int_{\prod_{j=1}^d [t_j,(t_j+1)/2)} \sum_{\boldsymbol{\delta}\in\{0,1\}^d} \Big(f\big((1-\delta_1)(t_1+1-x_1)+\delta_1 x_1, \dots, \\
&\qquad\qquad\qquad (1-\delta_d)(t_d+1-x_d)+\delta_d x_d\big)\Big)^2 \, d\mathbf{x} \\
&> \frac{C_d^2}{2^d} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^d r_j} \cdot \prod_{j=1}^d \int_{t_j}^{(t_j+1)/2} (t_j+1-2x_j)^2 \, dx_j \\
&= \frac{C_d^2}{12^d} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^d r_j} \cdot \prod_{j=1}^d (1-t_j)^3 \geq \frac{C_d^2}{12^d} \cdot t^2 = t^2,
\end{aligned}$$

which is a contradiction. Here, the last inequality follows from

$$1 - t_j \geq 1 - \frac{1}{2^{r_j}} \geq \frac{1}{2^{r_j}},$$

which holds since $r_j \geq 1$.

Next, suppose, for contradiction, that there exist $\mathbf{r} = (r_1, \dots, r_d) \in \mathbb{Z}_{\geq 0}^d$ and $1/2^{r_j+1} \leq s_j < t_j \leq 1/2^{r_j}$ for $j \in [d]$ such that

$$\begin{bmatrix} s_1, t_1 & \\ \vdots & ; \; f \\ s_d, t_d & \end{bmatrix} > C_d t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^d r_j},$$

where $C_d = (4\sqrt{6})^d =: C_{d,d}^U$. By Lemma 36, if $v_j < w_j \leq s_j$ for $j \in [d]$, then

$$\begin{bmatrix} v_1, w_1 & \\ \vdots & ; \; f \\ v_d, w_d & \end{bmatrix} \geq \begin{bmatrix} s_1, t_1 & \\ \vdots & ; \; f \\ s_d, t_d & \end{bmatrix} > C_d t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^d r_j}.$$

As in the previous case, applying Cauchy inequality gives

$$\sum_{\boldsymbol{\delta}\in\{0,1\}^d} \Big(f\big((1-\delta_1)w_1+\delta_1 v_1, \dots, (1-\delta_d)w_d+\delta_d v_d\big)\Big)^2 > \prod_{j=1}^d (w_j - v_j)^2 \cdot \frac{C_d^2}{2^d} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^d r_j}.$$

Since $x_j < s_j - x_j \leq s_j$ for all $x_j \in [0, s_j/2)$, it follows that

$$\begin{aligned}
t^2 &\geq \|f\|_2^2 \\
&\geq \int_{\prod_{j=1}^d [0,s_j/2)} \sum_{\boldsymbol{\delta}\in\{0,1\}^d} \Big(f\big((1-\delta_1)(s_1-x_1)+\delta_1 x_1, \dots, (1-\delta_d)(s_d-x_d)+\delta_d x_d\big)\Big)^2 \, d\mathbf{x}
\end{aligned}$$

$$
\begin{aligned}
&> \frac{C_d^2}{2^d} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^d r_j} \cdot \prod_{j=1}^d \int_0^{s_j/2} (s_j - 2x_j)^2 \, dx_j \\
&= \frac{C_d^2}{12^d} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^d r_j} \cdot \prod_{j=1}^d s_j^3 \geq \frac{C_d^2}{96^d} \cdot t^2 = t^2,
\end{aligned}
$$

which leads to contradiction. Here, the last inequality is due to the fact that $s_j \geq 1/2^{r_j+1}$.

Now, we assume that (D.38) and (D.39) hold with $C_d = C^L_{d,m+1}$ and $C_d = C^U_{d,m+1}$, respectively, for every subset $S \subseteq [d]$ with $|S| = m+1$, and prove that they also hold with some C_d for every subset $S \subseteq [d]$ with $|S| = m$. For notational convenience, we present the proof only for the case $S = [m] = \{1, \dots, m\}$; the other cases follow analogously.

Suppose, for contradiction, that there exist $\mathbf{r} = (r_1, \dots, r_d) \in \mathbb{N}^d$, $1/2^{r_j+1} \leq s_j < t_j \leq 1/2^{r_j}$ for $j = 1, \dots, m$, and $t_j \in [1/2^{r_j+1}, 1/2^{r_j}]$ for $j = m+1, \dots, d$ such that

$$
\begin{bmatrix} s_1, t_1 & \\ \vdots & \\ s_m, t_m & ; \ f \\ t_{m+1} & \\ \vdots & \\ t_d & \end{bmatrix} < -C_d t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^m r_j},
$$

where $C_d = (3/2)(d-m)C^U_{d,m+1} + (2\sqrt{3})^m =: C^L_{d,m}$. If $t_j \leq v_j < w_j$ for $j = 1, \dots, m$ and $t_j \leq v_j \leq 1/2^{r_j-1}$ for $j = m+1, \dots, d$, then we have

$$
\begin{aligned}
\begin{bmatrix} v_1, w_1 & \\ \vdots & \\ v_m, w_m & ; \ f \\ v_{m+1} & \\ \vdots & \\ v_d & \end{bmatrix}
&= \begin{bmatrix} v_1, w_1 & \\ \vdots & \\ v_m, w_m & \\ t_{m+1} & ; \ f \\ v_{m+2} & \\ \vdots & \\ v_d & \end{bmatrix}
+ (v_{m+1} - t_{m+1}) \begin{bmatrix} v_1, w_1 & \\ \vdots & \\ v_m, w_m & \\ t_{m+1}, v_{m+1} & ; \ f \\ v_{m+2} & \\ \vdots & \\ v_d & \end{bmatrix} \\
&\leq \begin{bmatrix} v_1, w_1 & \\ \vdots & \\ v_m, w_m & \\ t_{m+1} & ; \ f \\ v_{m+2} & \\ \vdots & \\ v_d & \end{bmatrix}
+ \frac{3}{2^{r_{m+1}+1}} \cdot C^U_{d,m+1} t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^{m+1} r_j},
\end{aligned}
$$

where the inequality follows from Lemma 36, the induction hypothesis, and the fact that

$$
v_{m+1} - t_{m+1} \leq \frac{1}{2^{r_{m+1}-1}} - \frac{1}{2^{r_{m+1}+1}} \leq \frac{3}{2^{r_{m+1}+1}}.
$$

Iterating this argument for the $(m+2)^{\text{th}}, \ldots, d^{\text{th}}$ coordinates yields

$$\begin{aligned}
\begin{bmatrix} v_1, w_1 & & \\ \vdots & & \\ v_m, w_m & ; & f \\ v_{m+1} & & \\ \vdots & & \\ v_d & & \end{bmatrix} &\leq \begin{bmatrix} v_1, w_1 & & \\ \vdots & & \\ v_m, w_m & ; & f \\ t_{m+1} & & \\ \vdots & & \\ t_d & & \end{bmatrix} + \frac{3}{2} \cdot (d-m) C^U_{d,m+1} t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^m r_j} \\
&\leq \begin{bmatrix} s_1, t_1 & & \\ \vdots & & \\ s_m, t_m & ; & f \\ t_{m+1} & & \\ \vdots & & \\ t_d & & \end{bmatrix} + \frac{3}{2} \cdot (d-m) C^U_{d,m+1} t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^m r_j} \\
&< -\Big(C_d - \frac{3}{2} \cdot (d-m) C^U_{d,m+1}\Big) t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^m r_j}.
\end{aligned}$$

By Cauchy inequality, it follows that

$$\begin{aligned}
&\sum_{\boldsymbol{\delta} \in \{0,1\}^m} \Big(f\big((1-\delta_1) w_1 + \delta_1 v_1, \ldots, (1-\delta_m) w_m + \delta_m v_m, v_{m+1}, \ldots, v_d\big)\Big)^2 \\
&\qquad > \prod_{j=1}^m (w_j - v_j)^2 \cdot \frac{\widetilde{C}_d^2}{2^m} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^m r_j},
\end{aligned}$$

where

$$\widetilde{C}_d := C_d - \frac{3}{2} \cdot (d-m) C^U_{d,m+1} = (2\sqrt{3})^m.$$

Hence, integrating as before, we obtain

$$\begin{aligned}
t^2 &\geq \|f\|_2^2 \\
&\geq \int_{\prod_{j=1}^m [t_j, (t_j+1)/2) \times \prod_{j=m+1}^d [t_j, 1/2^{r_j - 1}]} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} \\
&\qquad\qquad \Big(f\big((1-\delta_1)(t_1 + 1 - x_1) + \delta_1 x_1, \ldots, \\
&\qquad\qquad\qquad (1-\delta_m)(t_m + 1 - x_m) + \delta_m x_m, x_{m+1}, \ldots, x_d\big)\Big)^2 \, d\mathbf{x} \\
&> \frac{\widetilde{C}_d^2}{2^m} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^m r_j} \cdot \prod_{j=1}^m \int_{t_j}^{(t_j+1)/2} (t_j + 1 - 2x_j)^2 \, dx_j \cdot \prod_{j=m+1}^d \Big(\frac{1}{2^{r_j - 1}} - t_j\Big) \\
&\geq \frac{\widetilde{C}_d^2}{12^m} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^m r_j} \cdot \prod_{j=1}^m (1-t_j)^3 \cdot \prod_{j=m+1}^d \frac{1}{2^{r_j}} \geq \frac{\widetilde{C}_d^2}{12^m} \cdot t^2 = t^2,
\end{aligned}$$

which is a contradiction. Here, the second-to-last inequality uses the fact that

$$\frac{1}{2^{r_j-1}} - t_j \geq \frac{1}{2^{r_j-1}} - \frac{1}{2^{r_j}} = \frac{1}{2^{r_j}} \quad \text{for } j = m+1, \dots, d.$$

Next, suppose, for contradiction, that there exist $\mathbf{r} = (r_1, \dots, r_d) \in \mathbb{Z}^d_{\geq 0}$, $1/2^{r_j+1} \leq s_j < t_j \leq 1/2^{r_j}$ for $j = 1, \dots, m$, and $t_j \in [1/2^{r_j+1}, 1/2^{r_j}]$ for $j = m+1, \dots, d$ such that

$$\begin{bmatrix} s_1, t_1 & & \\ \vdots & & \\ s_m, t_m & ; & f \\ t_{m+1} & & \\ \vdots & & \\ t_d & & \end{bmatrix} > C_d t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^m r_j},$$

where $C_d = (3/2)(d-m)C^U_{d,m+1} + (2\sqrt{6})^m \cdot 2^d =: C^U_{d,m}$. For $v_j < w_j \leq s_j$ for $j = 1, \dots, m$ and $1/2^{r_j+2} \leq v_j \leq t_j$ for $j = m+1, \dots, d$, we have

$$\begin{aligned} \begin{bmatrix} v_1, w_1 & & \\ \vdots & & \\ v_m, w_m & ; & f \\ v_{m+1} & & \\ \vdots & & \\ v_d & & \end{bmatrix} &= \begin{bmatrix} v_1, w_1 & & \\ \vdots & & \\ v_m, w_m & & \\ t_{m+1} & ; & f \\ v_{m+2} & & \\ \vdots & & \\ v_d & & \end{bmatrix} - (t_{m+1} - v_{m+1}) \begin{bmatrix} v_1, w_1 & & \\ \vdots & & \\ v_m, w_m & & \\ v_{m+1}, t_{m+1} & ; & f \\ v_{m+2} & & \\ \vdots & & \\ v_d & & \end{bmatrix} \\ &\geq \begin{bmatrix} v_1, w_1 & & \\ \vdots & & \\ v_m, w_m & & \\ t_{m+1} & ; & f \\ v_{m+2} & & \\ \vdots & & \\ v_d & & \end{bmatrix} - \frac{3}{2^{r_{m+1}+2}} \cdot C^U_{d,m+1} t \cdot 2^{r_+/2} \cdot 2^{(\sum_{j=1}^{m+1} r_j)+1}, \end{aligned}$$

where the inequality is due to Lemma 36, the induction hypothesis, and the fact that

$$t_{m+1} - v_{m+1} \leq \frac{1}{2^{r_{m+1}}} - \frac{1}{2^{r_{m+1}+2}} \leq \frac{3}{2^{r_{m+1}+2}}.$$

Repeating this argument with the $(m+2)^{\text{th}}, \dots, d^{\text{th}}$ coordinates yields

$$\begin{aligned}
\begin{bmatrix} v_1, w_1 & & \\ \vdots & & \\ v_m, w_m & ; & f \\ v_{m+1} & & \\ \vdots & & \\ v_d & & \end{bmatrix} &\geq \begin{bmatrix} v_1, w_1 & & \\ \vdots & & \\ v_m, w_m & ; & f \\ t_{m+1} & & \\ \vdots & & \\ t_d & & \end{bmatrix} - \frac{3}{2} \cdot (d-m) C^U_{d,m+1} t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^m r_j} \\
&\geq \begin{bmatrix} s_1, t_1 & & \\ \vdots & & \\ s_m, t_m & ; & f \\ t_{m+1} & & \\ \vdots & & \\ t_d & & \end{bmatrix} - \frac{3}{2} \cdot (d-m) C^U_{d,m+1} t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^m r_j} \\
&> \Big(C_d - \frac{3}{2} \cdot (d-m) C^U_{d,m+1} \Big) t \cdot 2^{r_+/2} \cdot 2^{\sum_{j=1}^m r_j}.
\end{aligned}$$

If we let

$$\widetilde{C}_d := C_d - \frac{3}{2} \cdot (d-m) C^U_{d,m+1} = (2\sqrt{6})^m \cdot 2^d,$$

then we have

$$\begin{aligned}
t^2 &\geq \|f\|_2^2 \\
&\geq \int_{\prod_{j=1}^m [0, s_j/2) \times \prod_{j=m+1}^d [1/2^{r_j+2}, t_j]} \sum_{\boldsymbol{\delta} \in \{0,1\}^m} \\
&\qquad \Big(f\big((1-\delta_1)(s_1 - x_1) + \delta_1 x_1, \dots, (1-\delta_m)(s_m - x_m) + \delta_m x_m, x_{m+1}, \dots, x_d \big) \Big)^2 \, d\mathbf{x} \\
&> \frac{\widetilde{C}_d^2}{2^m} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^m r_j} \cdot \prod_{j=1}^m \int_0^{s_j/2} (s_j - 2x_j)^2 \, dx_j \cdot \prod_{j=m+1}^d \Big(t_j - \frac{1}{2^{r_j+2}} \Big) \\
&\geq \frac{\widetilde{C}_d^2}{12^m} \cdot t^2 \cdot 2^{r_+} \cdot 2^{2\sum_{j=1}^m r_j} \cdot \prod_{j=1}^m s_j^3 \cdot \prod_{j=m+1}^d \frac{1}{2^{r_j+2}} \geq \frac{\widetilde{C}_d^2}{24^m \cdot 4^d} \cdot t^2 = t^2,
\end{aligned}$$

which is a contradiction.

For these reasons, (D.38) and (D.39) hold with $C_d = C^L_{d,m}$ and $C_d = C^U_{d,m}$, respectively, for every subset $S \subseteq [d]$ with $|S| = m$. By induction, it follows that (D.38) and (D.39) hold for every $\emptyset \neq S \subseteq [d]$ with $C_d = \max(\max_{m \in [d]} C^L_{d,m}, \max_{m \in [d]} C^U_{d,m})$. □

D.15 Proof of Lemma 34

Proof of Lemma 34. The proof of this lemma closely follows that of Lemma 17 in Appendix C. We therefore only describe the modifications required at the beginning of their proof in order to establish Lemma 34.

Suppose $f \in D(V,t)$ is of the form (4.1). For notational convenience, write $c_0 = c_\emptyset$. Here, we prove that there exists a constant $C_d > 0$ such that

$$|c_T| \le C_d(V+t)$$

for every subset $T \subseteq [d]$ with $|T| \le s$. Once this bound is established, the remainder of the proof follows exactly as in the proof of Lemma 17.

Since $V_{\text{mars}}(f) \le V$, we have

$$t \ge \|f\|_2 \ge \left\{ \int_{[0,1]^d} \left(\sum_{0 \le |S| \le s} c_S \prod_{j \in S} x_j \right)^2 d\mathbf{x} \right\}^{1/2} - V,$$

which implies

$$\int_{[0,1]^d} \left(\sum_{0 \le |S| \le s} c_S \prod_{j \in S} x_j \right)^2 d\mathbf{x} \le (V+t)^2. \tag{D.49}$$

Now, fix $\delta \in (0,1)$, and define $A = \sum_{0 \le |S| \le s} c_S^2$. Restricting the integral (D.49) to $[0,\delta]^d$ and applying Cauchy inequality repeatedly, we obtain

$$\begin{aligned}(V+t)^2 &\ge \int_{[0,\delta]^d} \left(\sum_{0 \le |S| \le s} c_S \prod_{j \in S} x_j \right)^2 d\mathbf{x} \ge \frac{1}{2} c_\emptyset^2 \delta^d - \int_{[0,\delta]^d} \left(\sum_{0 < |S| \le s} c_S \prod_{j \in S} x_j \right)^2 d\mathbf{x} \\ &\ge \frac{1}{2} c_\emptyset^2 \delta^d - A \int_{[0,\delta]^d} \sum_{0 < |S| \le s} \prod_{j \in S} x_j^2 \, d\mathbf{x} = \frac{1}{2} c_\emptyset^2 \delta^d - C_d \delta^{d+2} A,\end{aligned}$$

and therefore,

$$c_\emptyset^2 \le 2(V+t)^2 \delta^{-d} + C_d \delta^2 A.$$

Next, assume that for every $T \subseteq [d]$ with $|T| < m \le s$, we have

$$c_T^2 \le C_d (V+t)^2 \delta^{-d} + C_d \delta^2 A. \tag{D.50}$$

We now show that the same bound, possibly with a different constant C_d, also holds for all subsets $T \subseteq [d]$ with $|T| = m$. Without loss of generality, it suffices to consider the case $T = [m] = \{1, \dots, m\}$, since the argument is identical for any other subset of size m.

Restricting the integral (D.49) to $[0,1]^m \times [0,\delta]^{d-m}$ and applying Cauchy inequality yields

$$(V+t)^2 \ge \int_{[0,1]^m \times [0,\delta]^{d-m}} \left(\sum_{0 \le |S| \le s} c_S \prod_{j \in S} x_j \right)^2 d\mathbf{x}$$

$$
\begin{aligned}
&\geq \frac{1}{3} c_{[m]}^2 \delta^{d-m} - \int_{[0,1]^m \times [0,\delta]^{d-m}} \left[\Big(\sum_{S \subsetneq [m]} c_S \prod_{j \in S} x_j \Big)^2 + \Big(\sum_{\substack{0 \leq |S| \leq s \\ S \not\subseteq [m]}} c_S \prod_{j \in S} x_j \Big)^2 \right] d\mathbf{x} \\
&\geq \frac{1}{3} c_{[m]}^2 \delta^{d-m} - \Big(\sum_{S \subsetneq [m]} c_S^2 \Big) \cdot \int_{[0,1]^m \times [0,\delta]^{d-m}} \sum_{S \subsetneq [m]} \prod_{j \in S} x_j^2 \, d\mathbf{x} \\
&\qquad - A \int_{[0,1]^m \times [0,\delta]^{d-m}} \sum_{\substack{0 \leq |S| \leq s \\ S \not\subseteq [m]}} \prod_{j \in S} x_j^2 \, d\mathbf{x}.
\end{aligned}
$$

By the induction hypothesis, we thus have

$$
\begin{aligned}
(V+t)^2 &\geq \frac{1}{3} c_{[m]}^2 \delta^{d-m} - \sum_{S \subsetneq [m]} \big(C_d (V+t)^2 \delta^{-d} + C_d \delta^2 A \big) \delta^{d-m} - C_d \delta^{d-m+2} A \\
&\geq \frac{1}{3} c_{[m]}^2 \delta^{d-m} - C_d (V+t)^2 \delta^{-m} - C_d \delta^{d-m+2} A.
\end{aligned}
$$

Rearranging terms, we obtain

$$
c_{[m]}^2 \leq 3(V+t)^2 \delta^{m-d} + C_d (V+t)^2 \delta^{-d} + C_d \delta^2 A \leq C_d (V+t)^2 \delta^{-d} + C_d \delta^2 A.
$$

This proves that the bound (D.50) holds for $T = [m]$, and hence, for all $T \subseteq [d]$ with $|T| = m$. By induction, it follows that the bound (D.50) holds for every $T \subseteq [d]$ with $|T| \leq s$.

Summing the bounds (D.50) over all subsets $T \subseteq [d]$ with $|T| \leq s$, we obtain

$$
A \leq C_d (V+t)^2 \delta^{-d} + C_d \delta^2 A.
$$

Hence,

$$
A \leq \frac{C_d (V+t)^2 \delta^{-d}}{1 - C_d \delta^2}
$$

provided $C_d \delta^2 < 1$. Choosing $\delta \in (0,1)$ sufficiently small and substituting this bound back into (D.50), we obtain

$$
c_T^2 \leq C_d (V+t)^2
$$

for all $T \subseteq [d]$ with $|T| \leq s$, as desired. $\square$

Appendix E

Proofs for XGBoost

E.1 Proof of Proposition 8

Proof of Proposition 8. We have already seen that any element of $\mathcal{F}_{\text{st}}^{d,s}$ admits a representation of the form (5.4) with discrete signed measures $\nu_{L,U}$ with finite support. It therefore suffices to prove the converse inclusion.

Observe that each atom $A_{\mathbf{l},\mathbf{u}}^{L,U}$ can be viewed as a regression tree with right-continuous splits and depth at most s, whose leaf weights are all zero except for a single leaf with weight one. Consequently, each atom $A_{\mathbf{l},\mathbf{u}}^{L,U}$ belongs to $\mathcal{F}_{\text{st}}^{d,s}$. Since $\mathcal{F}_{\text{st}}^{d,s}$ is closed under addition and scalar multiplication, it follows that any finite linear combination of these atoms—and hence any function of the form (5.4) with discrete signed measures $\nu_{L,U}$ having finite support—also belongs to $\mathcal{F}_{\text{st}}^{d,s}$. This completes the proof. $\square$

E.2 Proof of Proposition 10

Proof of Proposition 10. Recall that the sum in (5.4) ranges over all $L, U \subseteq [d]$ with $0 < |L| + |U| \leq s$. Hence, for each function f, the set of admissible representations $f_{c,\{\nu_{L,U}\}} \equiv f$ enlarges as s increases. Since $V_{\infty-\text{xgb}}^{d,s}(\cdot)$ is defined as an infimum over these representations, this gives (a).

Since $|L| + |U|$ is always bounded by $2d$, increasing s beyond $2d$ does not enlarge the set of admissible representations. Hence, $V_{\infty-\text{xgb}}^{d,s}(\cdot)$ stabilizes once $s \geq 2d$, which proves (b).

Lastly, (c) follows from the convexity of total variation $\|\cdot\|_{\text{TV}}$ on the space of finite signed measures. $\square$

E.3 Proof of Theorem 14

Before proving the theorem, we first observe and prove the following alternative characterization of $V_{\text{xgb}}^{d,s}(\cdot)$, originally defined via (5.1).

Lemma 37. *The complexity measure $V_{xgb}^{d,s}(\cdot)$ can be alternatively characterized as*

$$V_{xgb}^{d,s}(f) = \inf \Bigg\{ \sum_{0<|L|+|U|\le s} \|\nu_{L,U}\|_{TV} : f_{c,\{\nu_{L,U}\}} \equiv f \text{ and } \nu_{L,U} \text{ are discrete signed measures with finite support} \Bigg\}.$$

Note that the only difference from the definition (5.5) of $V_{\infty-\mathrm{xgb}}^{d,s}(\cdot)$ is that the signed measures $\nu_{L,U}$ are restricted to be discrete with finite support. We will show in the proof of Theorem 14 that this additional restriction does not affect the value of the infimum for functions in $\mathcal{F}_{\mathrm{st}}^{d,s}$.

Proof of Lemma 37. Suppose first that all $\nu_{L,U}$ are discrete signed measures with finite support. Then, $f_{c,\{\nu_{L,U}\}}$ is a finite linear combination of the atoms $A_{\mathbf{l},\mathbf{u}}^{L,U}$ with coefficients $\nu_{L,U}(\{(\mathbf{l},\mathbf{u})\})$ (plus a constant). Recall that each atom $A_{\mathbf{l},\mathbf{u}}^{L,U}$ can be viewed as a regression tree with right-continuous splits and depth at most s, whose leaf weights are all zero except for a single leaf with weight one. Consequently, $f_{c,\{\nu_{L,U}\}}$ can be represented as a finite sum of regression trees of the same type, whose leaf weight vectors each contain a single nonzero entry equal to $\nu_{L,U}(\{(\mathbf{l},\mathbf{u})\})$. For this representation, the total ℓ^1 norm of the leaf weight vectors is exactly equal to the sum of the total variations of $\nu_{L,U}$. This proves that the infimum in the lemma is greater than or equal to the infimum in (5.1).

Now, suppose that $f \in \mathcal{F}_{\mathrm{st}}^{d,s}$ and that it is represented as a finite sum of regression trees with right-continuous splits and depth at most s. Let $\mathbf{w}_k$ denote the leaf weight vector of the k^{th} tree. By decomposing each tree into the atoms $A_{\mathbf{l},\mathbf{u}}^{L,U}$ corresponding to paths from the root to each leaf, we obtain a representation of f as a finite linear combination of these atoms whose coefficient vector has ℓ^1 norm no larger than $\sum_k \|\mathbf{w}_k\|_1$. Equivalently, there exists a representation $f_{c,\{\nu_{L,U}\}} \equiv f$ with discrete signed measures $\nu_{L,U}$ having finite support such that the sum of the total variations of $\nu_{L,U}$ is no larger than $\sum_k \|\mathbf{w}_k\|_1$. This shows that the infimum in the lemma is no larger than the infimum in (5.1). $\square$

Proof of Theorem 14. We begin by introducing some notation used in the proof. For a function $g : \mathbb{R}^d \to \mathbb{R}$ and a nonempty subset $S \subseteq [d]$, define

$$g^S(x_j, j \in S) := \lim_{(x_j, j\in S^c)\to(-\infty, j\in S^c)} g(x_1, \dots, x_d) \qquad \text{for } (x_j, j \in S) \in \mathbb{R}^{|S|}$$

whenever the limit exists, where $S^c = [d] \setminus S$.

Fix $f \in \mathcal{F}_{\mathrm{st}}^{d,s}$. Since f is a finite sum of regression trees, there exists a partition $-\infty = v_0^{(j)} < v_1^{(j)} < \cdots < v_{n_j}^{(j)} < v_{n_j+1}^{(j)} = +\infty$ of $\mathbb{R}$ for each $j \in [d]$ such that f is constant on

$$\prod_{j\in S}(v_{m_j}^{(j)}, v_{m_j+1}^{(j)}) \times \prod_{j\in S^c}\{v_{m_j}^{(j)}\} \tag{E.1}$$

for every nonempty $S \subseteq [d]$, $m_j \in \{0, \dots, n_j\}$ for $j \in S$, and $m_j \in \{1, \dots, n_j\}$ for $j \in S^c$.

For each nonempty $S \subseteq [d]$ with $|S| \le s$ and $\mathbf{m} = (m_j, j \in S) \in \prod_{j \in S}[n_j]$, define the alternating-sum functional

$$\Delta^S_{\mathbf{m}}(g) := \lim_{\epsilon \to 0+} \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S} \delta_j} \cdot g^S\big(v^{(j)}_{m_j} - \delta_j \epsilon, j \in S\big)$$

for piecewise constant functions g as in (E.1).

Suppose $f_{c,\{\nu_{L,U}\}} \equiv f$. Then, clearly, we have

$$\Delta^S_{\mathbf{m}}(f_{c,\{\nu_{L,U}\}}) = \Delta^S_{\mathbf{m}}(f) \tag{E.2}$$

for all nonempty $S \subseteq [d]$ with $|S| \le s$ and $\mathbf{m} \in \prod_{j \in S}[n_j]$. In fact, condition (E.2) captures almost all of the information contained in the identity $f_{c,\{\nu_{L,U}\}} \equiv f$. The following lemma, whose proof is given after the current proof, makes this precise. This lemma will play an important role later.

Lemma 38. *If $g, h \in \mathcal{F}^{d,s}_{st}$ are piecewise constant as in* (E.1) *and satisfy*

$$\Delta^S_{\mathbf{m}}(g) = \Delta^S_{\mathbf{m}}(h)$$

for all nonempty $S \subseteq [d]$ with $|S| \le s$ and $\mathbf{m} \in \prod_{j \in S}[n_j]$, then g and h differ only by an additive constant; that is, there exists $b \in \mathbb{R}$ such that $g(\mathbf{x}) = h(\mathbf{x}) + b$ for all $\mathbf{x} \in \mathbb{R}^d$.

We simplify (E.2) and express it in terms of $\nu_{L,U}$ more explicitly. Fix a nonempty $S \subseteq [d]$ with $|S| \le s$ and $\mathbf{m} = (m_j, j \in S) \in \prod_{j \in S}[n_j]$. Expanding the left-hand side of (E.2) gives

$$\begin{aligned}\Delta^S_{\mathbf{m}}(f_{c,\{\nu_{L,U}\}}) = &\sum_{\substack{L,U : L \subseteq S \subseteq L \cup U \\ |L|+|U| \le s}} \lim_{\epsilon \to 0+} \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S} \delta_j} \\ &\cdot \int_{\mathbb{R}^{|L|+|U|}} \prod_{j \in L} \mathbf{1}\big(v^{(j)}_{m_j} - \delta_j \epsilon \ge l_j\big) \cdot \prod_{j \in U \cap S} \mathbf{1}\big(v^{(j)}_{m_j} - \delta_j \epsilon < u_j\big)\, d\nu_{L,U}(\mathbf{l}, \mathbf{u}).\end{aligned} \tag{E.3}$$

The inner limit can be simplified by exchanging the order of summation and integration and analyzing each indicator term. Specifically,

$$\begin{aligned}&\lim_{\epsilon \to 0+} \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S} \delta_j} \int_{\mathbb{R}^{|L|+|U|}} \prod_{j \in L} \mathbf{1}\big(v^{(j)}_{m_j} - \delta_j \epsilon \ge l_j\big) \cdot \prod_{j \in U \cap S} \mathbf{1}\big(v^{(j)}_{m_j} - \delta_j \epsilon < u_j\big)\, d\nu_{L,U}(\mathbf{l}, \mathbf{u}) \\ &= \lim_{\epsilon \to 0+} \int_{\mathbb{R}^{|L|+|U|}} \prod_{j \in L \setminus U} \big\{\mathbf{1}\big(v^{(j)}_{m_j} \ge l_j\big) - \mathbf{1}\big(v^{(j)}_{m_j} - \epsilon \ge l_j\big)\big\} \\ &\qquad \cdot \prod_{j \in L \cap U} \big\{\mathbf{1}\big(l_j \le v^{(j)}_{m_j} < u_j\big) - \mathbf{1}\big(l_j \le v^{(j)}_{m_j} - \epsilon < u_j\big)\big\} \\ &\qquad \cdot \prod_{j \in (U \setminus L) \cap S} \big\{\mathbf{1}\big(v^{(j)}_{m_j} < u_j\big) - \mathbf{1}\big(v^{(j)}_{m_j} - \epsilon < u_j\big)\big\}\, d\nu_{L,U}(\mathbf{l}, \mathbf{u})\end{aligned}$$

$$= (-1)^{|(U\setminus L)\cap S|} \sum_{K\subseteq L\cap U} (-1)^{|(L\cap U)\setminus K|} \cdot \nu_{L,U}\Big(\prod_{j\in(L\setminus U)\cup K} \{v^{(j)}_{m_j}\} \times \prod_{j\in(L\cap U)\setminus K} (-\infty, v^{(j)}_{m_j}) \times \prod_{j\in K} (v^{(j)}_{m_j}, +\infty) \times \prod_{\substack{j\in((L\cap U)\setminus K)\\ \cup((U\setminus L)\cap S)}} \{v^{(j)}_{m_j}\} \times \mathbb{R}^{|U\setminus S|}\Big).$$

Thus, condition (E.2), which holds for all nonempty $S\subseteq[d]$ with $|S|\le s$ and $\mathbf{m}=(m_j, j\in S)\in\prod_{j\in S}[n_j]$ provided that $f_{c,\{\nu_{L,U}\}}\equiv f$, can be written as

$$\sum_{\substack{L,U: L\subseteq S\subseteq L\cup U\\ |L|+|U|\le s}} (-1)^{|(U\setminus L)\cap S|} \sum_{K\subseteq L\cap U} (-1)^{|(L\cap U)\setminus K|} \cdot \nu_{L,U}\Big(\prod_{j\in(L\setminus U)\cup K} \{v^{(j)}_{m_j}\} \times \prod_{j\in(L\cap U)\setminus K} (-\infty, v^{(j)}_{m_j}) \times \prod_{j\in K} (v^{(j)}_{m_j}, +\infty) \times \prod_{\substack{j\in((L\cap U)\setminus K)\\ \cup((U\setminus L)\cap S)}} \{v^{(j)}_{m_j}\} \times \mathbb{R}^{|U\setminus S|}\Big) = \Delta^S_{\mathbf{m}}(f). \tag{E.4}$$

Consequently, we have

$$V^{d,s}_{\infty-\mathrm{xgb}}(f) \ge \inf\Big\{ \sum_{0<|L|+|U|\le s} \|\nu_{L,U}\|_{\mathrm{TV}} : \nu_{L,U} \text{ satisfy (E.4)} \Big\}. \tag{E.5}$$

Now, we show the infimum in (E.5) is achieved by discrete signed measures supported on

$$\prod_{j\in L}\{v^{(j)}_1,\dots,v^{(j)}_{n_j}\} \times \prod_{j\in U}\{v^{(j)}_1,\dots,v^{(j)}_{n_j}\}. \tag{E.6}$$

Suppose $\nu_{L,U}$ are signed measures satisfying (E.4). For each $j\in[d]$, let $V^{(j)}=\{v^{(j)}_1,\dots,v^{(j)}_{n_j}\}$, $\overline{V}^{(j)}_{m_j}=\{v^{(j)}_{m_j+1},\dots,v^{(j)}_{n_j}\}$, and $\underline{V}^{(j)}_{m_j}=\{v^{(j)}_1,\dots,v^{(j)}_{m_j-1}\}$. Define discrete signed measures $\mu_{L,U}$, supported on the lattices (E.6), by

$$\mu_{L,U}\big(\{(v^{(j)}_{p_j}, j\in L; v^{(j)}_{q_j}, j\in U)\}\big) = \sum_{\substack{\widetilde{U}\supseteq U\\ |\widetilde{U}|\le s-|L|}} \sum_{T\subseteq(L\cap\widetilde{U})\setminus U} (-1)^{|((L\cap\widetilde{U})\setminus U)\setminus T|} \cdot \nu_{L,\widetilde{U}}\Big(\prod_{j\in L\setminus(((L\cap\widetilde{U})\setminus U)\setminus T)} \{v^{(j)}_{p_j}\} \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} \big((-\infty, v^{(j)}_{p_j})\setminus \underline{V}^{(j)}_{p_j}\big) \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} \{v^{(j)}_{p_j}\} \times \prod_{j\in T} \big((v^{(j)}_{p_j}, +\infty)\setminus \overline{V}^{(j)}_{p_j}\big) \times \prod_{j\in U}\{v^{(j)}_{q_j}\} \times \prod_{j\in(\widetilde{U}\setminus L)\setminus U} (\mathbb{R}\setminus V^{(j)})\Big)$$

for $(p_j, j\in L)\in\prod_{j\in L}[n_j]$ and $(q_j, j\in U)\in\prod_{j\in U}[n_j]$. Observe that for $L\subseteq S\subseteq L\cup U$ with $|L|+|U|\le s$,

$$\mu_{L,U}\Big(\prod_{j\in(L\setminus U)\cup K} \{v^{(j)}_{m_j}\} \times \prod_{j\in(L\cap U)\setminus K} (-\infty, v^{(j)}_{m_j}) \times \prod_{j\in K} (v^{(j)}_{m_j}, +\infty) \times \prod_{\substack{j\in((L\cap U)\setminus K)\\ \cup((U\setminus L)\cap S)}} \{v^{(j)}_{m_j}\} \times \mathbb{R}^{|U\setminus S|}\Big)$$

$$= \sum_{\mathbf{r} \in \prod_{j \in (L\cap U)\setminus K} \underline{V}^{(j)}_{m_j} \times \prod_{j\in K} \overline{V}^{(j)}_{m_j} \times \prod_{U\setminus S} V^{(j)}} \mu_{L,U}\Big(\prod_{j\in(L\setminus U)\cup K} \{v^{(j)}_{m_j}\} \times \prod_{j\in(L\cap U)\setminus K} \{v^{(j)}_{r_j}\}$$

$$\times \prod_{j\in K}\{v^{(j)}_{r_j}\} \times \prod_{\substack{j\in((L\cap U)\setminus K)\\ \cup((U\setminus L)\cap S)}} \{v^{(j)}_{m_j}\} \times \prod_{j\in U\setminus S} \{v^{(j)}_{r_j}\}\Big)$$

$$= \sum_{\mathbf{r} \in \prod_{j \in (L\cap U)\setminus K} \underline{V}^{(j)}_{m_j} \times \prod_{j\in K} \overline{V}^{(j)}_{m_j} \times \prod_{U\setminus S} V^{(j)}} \sum_{\substack{\widetilde{U}\supseteq U\\ |\widetilde{U}|\le s-|L|}} \sum_{T\subseteq (L\cap\widetilde{U})\setminus U} (-1)^{|((L\cap\widetilde{U})\setminus U)\setminus T|}$$

$$\cdot \nu_{L,\widetilde{U}}\Big(\prod_{j\in(L\setminus\widetilde{U})\cup K\cup T} \{v^{(j)}_{m_j}\} \times \prod_{j\in(L\cap U)\setminus K} \{v^{(j)}_{r_j}\} \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} \big((-\infty, v^{(j)}_{m_j}) \setminus \underline{V}^{(j)}_{m_j}\big)$$

$$\times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} \{v^{(j)}_{m_j}\} \times \prod_{j\in T}\big((v^{(j)}_{m_j}, +\infty) \setminus \overline{V}^{(j)}_{m_j}\big)$$

$$\times \prod_{j\in K}\{v^{(j)}_{r_j}\} \times \prod_{\substack{j\in((L\cap U)\setminus K)\\ \cup((U\setminus L)\cap S)}} \{v^{(j)}_{m_j}\} \times \prod_{j\in U\setminus S}\{v^{(j)}_{r_j}\} \times \prod_{j\in(\widetilde{U}\setminus L)\setminus U} (\mathbb{R}\setminus V^{(j)})\Big)$$

$$= \sum_{\substack{\widetilde{U}\supseteq U\\ |\widetilde{U}|\le s-|L|}} \sum_{T\subseteq (L\cap\widetilde{U})\setminus U} (-1)^{|((L\cap\widetilde{U})\setminus U)\setminus T|}$$

$$\cdot \nu_{L,\widetilde{U}}\Big(\prod_{j\in(L\setminus\widetilde{U})\cup K\cup T} \{v^{(j)}_{m_j}\} \times \prod_{j\in(L\cap U)\setminus K} \underline{V}^{(j)}_{m_j} \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} \big((-\infty, v^{(j)}_{m_j}) \setminus \underline{V}^{(j)}_{m_j}\big)$$

$$\times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} \{v^{(j)}_{m_j}\} \times \prod_{j\in T}\big((v^{(j)}_{m_j}, +\infty) \setminus \overline{V}^{(j)}_{m_j}\big)$$

$$\times \prod_{j\in K}\overline{V}^{(j)}_{m_j} \times \prod_{\substack{j\in((L\cap U)\setminus K)\\ \cup((U\setminus L)\cap S)}} \{v^{(j)}_{m_j}\} \times \prod_{j\in U\setminus S} V^{(j)} \times \prod_{j\in(\widetilde{U}\setminus L)\setminus U} (\mathbb{R}\setminus V^{(j)})\Big).$$

Hence, the left-hand side of (E.4) for $\mu_{L,U}$ becomes

$$\sum_{\substack{L,U: L\subseteq S\subseteq L\cup U\\ |L|+|U|\le s}} (-1)^{|(U\setminus L)\cap S|} \sum_{K\subseteq L\cap U} (-1)^{|(L\cap U)\setminus K|}$$

$$\cdot \mu_{L,U}\Big(\prod_{j\in(L\setminus U)\cup K} \{v^{(j)}_{m_j}\} \times \prod_{j\in(L\cap U)\setminus K} (-\infty, v^{(j)}_{m_j})$$

$$\times \prod_{j\in K}(v^{(j)}_{m_j}, +\infty) \times \prod_{\substack{j\in((L\cap U)\setminus K)\\ \cup((U\setminus L)\cap S)}} \{v^{(j)}_{m_j}\} \times \mathbb{R}^{|U\setminus S|}\Big)$$

$$= \sum_{\substack{L,U: L\subseteq S\subseteq L\cup U\\ |L|+|U|\le s}} (-1)^{|(U\setminus L)\cap S|} \sum_{K\subseteq L\cap U} (-1)^{|(L\cap U)\setminus K|} \sum_{\substack{\widetilde{U}\supseteq U\\ |\widetilde{U}|\le s-|L|}} \sum_{T\subseteq (L\cap\widetilde{U})\setminus U} (-1)^{|((L\cap\widetilde{U})\setminus U)\setminus T|}$$

$$\cdot\, \nu_{L,\widetilde{U}}\Big(\prod_{j \in (L\setminus \widetilde{U})\cup K \cup T} \{v_{m_j}^{(j)}\} \times \prod_{j\in (L\cap U)\setminus K} \underline{V}_{m_j}^{(j)} \times \prod_{j \in ((L\cap \widetilde{U})\setminus U)\setminus T} \big((-\infty, v_{m_j}^{(j)}) \setminus \underline{V}_{m_j}^{(j)}\big)$$
$$\times \prod_{j \in ((L\cap \widetilde{U})\setminus U)\setminus T} \{v_{m_j}^{(j)}\} \times \prod_{j\in T} \big((v_{m_j}^{(j)}, +\infty) \setminus \overline{V}_{m_j}^{(j)}\big)$$
$$\times \prod_{j\in K} \overline{V}_{m_j}^{(j)} \times \prod_{\substack{j \in ((L\cap U)\setminus K) \\ \cup((U\setminus L)\cap S)}} \{v_{m_j}^{(j)}\} \times \prod_{j\in U\setminus S} V^{(j)} \times \prod_{j\in (\widetilde{U}\setminus L)\setminus U} (\mathbb{R}\setminus V^{(j)})\Big).$$

By changing the order of summation in U and $\widetilde{U}$ and combining the summations in K and T via $\widetilde{K} = K \cup T$, we can simplify the right-hand side to

$$\sum_{\substack{L,\widetilde{U}: L\subseteq S\subseteq L\cup \widetilde{U} \\ |L|+|\widetilde{U}|\le s}} (-1)^{|(\widetilde{U}\setminus L)\cap S|} \sum_{U: S\setminus L\subseteq U\subseteq \widetilde{U}} \sum_{K\subseteq L\cap U} \sum_{T\subseteq (L\cap\widetilde{U})\setminus U} (-1)^{|(L\cap U)\setminus K|} \cdot (-1)^{|((L\cap \widetilde{U})\setminus U)\setminus T|}$$
$$\cdot\, \nu_{L,\widetilde{U}}\Big(\prod_{j \in (L\setminus \widetilde{U})\cup K \cup T} \{v_{m_j}^{(j)}\} \times \prod_{j\in (L\cap U)\setminus K} \underline{V}_{m_j}^{(j)} \times \prod_{j \in ((L\cap \widetilde{U})\setminus U)\setminus T} \big((-\infty, v_{m_j}^{(j)}) \setminus \underline{V}_{m_j}^{(j)}\big)$$
$$\times \prod_{j \in ((L\cap \widetilde{U})\setminus U)\setminus T} \{v_{m_j}^{(j)}\} \times \prod_{j\in T} \big((v_{m_j}^{(j)}, +\infty) \setminus \overline{V}_{m_j}^{(j)}\big)$$
$$\times \prod_{j\in K} \overline{V}_{m_j}^{(j)} \times \prod_{\substack{j \in ((L\cap U)\setminus K) \\ \cup((U\setminus L)\cap S)}} \{v_{m_j}^{(j)}\} \times \prod_{j\in U\setminus S} V^{(j)} \times \prod_{j\in (\widetilde{U}\setminus L)\setminus U} (\mathbb{R}\setminus V^{(j)})\Big)$$
$$= \sum_{\substack{L,\widetilde{U}: L\subseteq S\subseteq L\cup \widetilde{U} \\ |L|+|\widetilde{U}|\le s}} (-1)^{|(\widetilde{U}\setminus L)\cap S|} \sum_{\widetilde{K}\subseteq L\cap\widetilde{U}} (-1)^{|(L\cap\widetilde{U})\setminus\widetilde{K}|} \sum_{U: S\setminus L\subseteq U\subseteq \widetilde{U}}$$
$$\cdot\, \nu_{L,\widetilde{U}}\Big(\prod_{j \in (L\setminus \widetilde{U})\cup \widetilde{K}} \{v_{m_j}^{(j)}\} \times \prod_{j\in ((L\cap\widetilde{U})\setminus\widetilde{K})\cap U} \underline{V}_{m_j}^{(j)} \times \prod_{j \in ((L\cap \widetilde{U})\setminus \widetilde{K})\setminus U} \big((-\infty, v_{m_j}^{(j)}) \setminus \underline{V}_{m_j}^{(j)}\big)$$
$$\times \prod_{j\in \widetilde{K}\setminus U} \big((v_{m_j}^{(j)}, +\infty) \setminus \overline{V}_{m_j}^{(j)}\big) \times \prod_{j\in\widetilde{K}\cap U} \overline{V}_{m_j}^{(j)} \times \prod_{\substack{j \in ((L\cap \widetilde{U})\setminus \widetilde{K}) \\ \cup((\widetilde{U}\setminus L)\cap S)}} \{v_{m_j}^{(j)}\}$$
$$\times \prod_{j\in U\setminus S} V^{(j)} \times \prod_{j\in (\widetilde{U}\setminus L)\setminus U} (\mathbb{R}\setminus V^{(j)})\Big).$$

Computing the inner summation over U yields

$$\sum_{\substack{L,U: L\subseteq S\subseteq L\cup U \\ |L|+|U|\le s}} (-1)^{|(U\setminus L)\cap S|} \sum_{K\subseteq L\cap U} (-1)^{|(L\cap U)\setminus K|}$$
$$\cdot\, \mu_{L,U}\Big(\prod_{j\in (L\setminus U)\cup K} \{v_{m_j}^{(j)}\} \times \prod_{j\in (L\cap U)\setminus K} (-\infty, v_{m_j}^{(j)})$$

$$\times \prod_{j\in K}(v_{m_j}^{(j)},+\infty) \times \prod_{\substack{j\in((L\cap U)\setminus K)\\ \cup((U\setminus L)\cap S)}} \{v_{m_j}^{(j)}\} \times \mathbb{R}^{|U\setminus S|}\Big)$$

$$= \sum_{\substack{L,\widetilde{U}:L\subseteq S\subseteq L\cup\widetilde{U}\\ |L|+|\widetilde{U}|\le s}} (-1)^{|(\widetilde{U}\setminus L)\cap S|} \sum_{\widetilde{K}\subseteq L\cap\widetilde{U}} (-1)^{|(L\cap\widetilde{U})\setminus\widetilde{K}|}$$

$$\cdot\, \nu_{L,\widetilde{U}}\Big(\prod_{j\in(L\setminus\widetilde{U})\cup\widetilde{K}}\{v_{m_j}^{(j)}\} \times \prod_{j\in((L\cap\widetilde{U})\setminus\widetilde{K})}(-\infty,v_{m_j}^{(j)})$$

$$\times \prod_{j\in\widetilde{K}}(v_{m_j}^{(j)},+\infty) \times \prod_{\substack{j\in((L\cap\widetilde{U})\setminus\widetilde{K})\\ \cup((\widetilde{U}\setminus L)\cap S)}} \{v_{m_j}^{(j)}\} \times \mathbb{R}^{|\widetilde{U}\setminus S|}\Big) = \Delta_{\mathbf{m}}^S(f),$$

which shows that (E.4) also holds for $\mu_{L,U}$.

Moreover, by the definition of $\mu_{L,U}$, we have

$$\sum_{0<|L|+|U|\le s} |\mu_{L,U}|(\mathbb{R}^{|L|+|U|})$$

$$= \sum_{0<|L|+|U|\le s} \sum_{\mathbf{p}\in\prod_{j\in L}[n_j]} \sum_{\mathbf{q}\in\prod_{j\in U}[n_j]} \big|\mu_{L,U}\big(\{(v_{p_j}^{(j)}, j\in L)\times(v_{q_j}^{(j)}, j\in U)\}\big)\big|$$

$$\le \sum_{0<|L|+|U|\le s} \sum_{\mathbf{p}\in\prod_{j\in L}[n_j]} \sum_{\mathbf{q}\in\prod_{j\in U}[n_j]} \sum_{\substack{\widetilde{U}\supseteq U\\ |\widetilde{U}|\le s-|L|}} \sum_{T\subseteq(L\cap\widetilde{U})\setminus U}$$

$$|\nu_{L,\widetilde{U}}|\Big(\prod_{j\in L\setminus(((L\cap\widetilde{U})\setminus U)\setminus T)}\{v_{p_j}^{(j)}\} \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T}\big((-\infty,v_{p_j}^{(j)})\setminus \underline{V}_{p_j}^{(j)}\big)$$

$$\times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T}\{v_{p_j}^{(j)}\} \times \prod_{j\in T}\big((v_{p_j}^{(j)},+\infty)\setminus \overline{V}_{p_j}^{(j)}\big)$$

$$\times \prod_{j\in U}\{v_{q_j}^{(j)}\} \times \prod_{j\in(\widetilde{U}\setminus L)\setminus U}(\mathbb{R}\setminus V^{(j)})\Big)$$

$$= \sum_{0<|L|+|\widetilde{U}|\le s} \sum_{U\subseteq\widetilde{U}} \sum_{T\subseteq(L\cap\widetilde{U})\setminus U} \sum_{\mathbf{p}\in\prod_{j\in L}[n_j]} \sum_{\mathbf{q}\in\prod_{j\in U}[n_j]}$$

$$|\nu_{L,\widetilde{U}}|\Big(\prod_{j\in L\setminus(((L\cap\widetilde{U})\setminus U)\setminus T)}\{v_{p_j}^{(j)}\} \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T}\big((-\infty,v_{p_j}^{(j)})\setminus \underline{V}_{p_j}^{(j)}\big)$$

$$\times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T}\{v_{p_j}^{(j)}\} \times \prod_{j\in T}\big((v_{p_j}^{(j)},+\infty)\setminus \overline{V}_{p_j}^{(j)}\big)$$

$$\times \prod_{j\in U}\{v_{q_j}^{(j)}\} \times \prod_{j\in(\widetilde{U}\setminus L)\setminus U}(\mathbb{R}\setminus V^{(j)})\Big)$$

$$
\begin{aligned}
\leq & \sum_{0<|L|+|\widetilde{U}|\leq s} \sum_{U\subseteq\widetilde{U}} \sum_{T\subseteq(L\cap\widetilde{U})\setminus U} \sum_{\mathbf{p}\in\prod_{j\in L}[n_j]} \sum_{\mathbf{q}\in\prod_{j\in U}[n_j]} \\
& \quad |\nu_{L,\widetilde{U}}|\Big(\prod_{j\in L\setminus(((L\cap\widetilde{U})\setminus U)\setminus T)} \{v^{(j)}_{p_j}\} \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} (\mathbb{R}\setminus V^{(j)}) \\
& \qquad \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} \{v^{(j)}_{p_j}\} \times \prod_{j\in T}(\mathbb{R}\setminus V^{(j)}) \times \prod_{j\in U}\{v^{(j)}_{q_j}\} \times \prod_{j\in(\widetilde{U}\setminus L)\setminus U} (\mathbb{R}\setminus V^{(j)})\Big) \\
= & \sum_{0<|L|+|\widetilde{U}|\leq s} \sum_{U\subseteq\widetilde{U}} \sum_{T\subseteq(L\cap\widetilde{U})\setminus U} \\
& \quad |\nu_{L,\widetilde{U}}|\Big(\prod_{j\in L\setminus(((L\cap\widetilde{U})\setminus U)\setminus T)} V^{(j)} \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} (\mathbb{R}\setminus V^{(j)}) \\
& \qquad \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} V^{(j)} \times \prod_{j\in T}(\mathbb{R}\setminus V^{(j)}) \times \prod_{j\in U} V^{(j)} \times \prod_{j\in(\widetilde{U}\setminus L)\setminus U} (\mathbb{R}\setminus V^{(j)})\Big) \\
\leq & \sum_{0<|L|+|\widetilde{U}|\leq s} |\nu_{L,\widetilde{U}}|(\mathbb{R}^{|L|+|\widetilde{U}|}).
\end{aligned}
$$

Here, we change the order of summation in U and $\widetilde{U}$ for the second equality, and for the second inequality, we use the fact that

$$
(-\infty, v^{(j)}_{p_j}) \setminus \underline{V}^{(j)}_{p_j} \subseteq \mathbb{R}\setminus V^{(j)} \text{ and } (v^{(j)}_{p_j}, +\infty) \setminus \overline{V}^{(j)}_{p_j} \subseteq \mathbb{R}\setminus V^{(j)} \text{ for all } j.
$$

The last inequality follows from the observation that for each L and $\widetilde{U}$, the sets

$$
\begin{aligned}
& \prod_{j\in L\setminus(((L\cap\widetilde{U})\setminus U)\setminus T)} V^{(j)} \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} (\mathbb{R}\setminus V^{(j)}) \\
& \quad \times \prod_{j\in((L\cap\widetilde{U})\setminus U)\setminus T} V^{(j)} \times \prod_{j\in T}(\mathbb{R}\setminus V^{(j)}) \times \prod_{j\in U} V^{(j)} \times \prod_{j\in(\widetilde{U}\setminus L)\setminus U} (\mathbb{R}\setminus V^{(j)})
\end{aligned}
$$

are pairwise disjoint as U ranges over subsets of $\widetilde{U}$ and T ranges over subsets of $(L\cap\widetilde{U})\setminus U$. This shows that the objective function of (E.5) for $\mu_{L,U}$ is no larger than that for $\nu_{L,U}$ and ensures that the infimum of (E.5) is attained by some discrete signed measures $\mu_{L,U}$ supported on (E.6).

Suppose $\mu_{L,U}$ are discrete signed measures supported on the lattices (E.6) that satisfy (E.4). We can parametrize these measures by

$$
\mu_{L,U}\big(\{(v^{(j)}_{p_j}, j\in L) \times (v^{(j)}_{q_j}, j\in U)\}\big) = \beta^{L,U}_{\mathbf{p},\mathbf{q}} \tag{E.7}
$$

for $L, U \subseteq [d]$ with $0 < |L| + |U| \leq s$, $\mathbf{p} = (p_j, j\in L) \in \prod_{j\in L}[n_j]$, and $\mathbf{q} = (q_j, j\in U) \in \prod_{j\in U}[n_j]$. Under this parametrization, (E.4) can be written entirely in terms of $\beta^{L,U}_{\mathbf{p},\mathbf{q}}$ as

$$
\sum_{\substack{L,U: L\subseteq S\subseteq L\cup U \\ |L|+|U|\leq s}} (-1)^{|(U\setminus L)\cap S|} \sum_{K\subseteq L\cap U} (-1)^{|(L\cap U)\setminus K|} \sum_{\mathbf{r}\in\prod_{j\in(L\cap U)\setminus K} \underline{V}^{(j)}_{m_j} \times \prod_{j\in K} \overline{V}^{(j)}_{m_j} \times \prod_{U\setminus S} V^{(j)}}
$$

$$\beta^{L,U}_{(m_j, j\in(L\setminus U)\cup K; r_j, j\in(L\cap U)\setminus K)\times(r_j, j\in K; m_j, j\in((L\cap U)\setminus K)\cup((U\setminus L)\cap S); r_j, j\in U\setminus S)} = \Delta^S_{\mathbf{m}}(f). \quad \text{(E.8)}$$

Consequently, (E.5) becomes

$$V^{d,s}_{\infty-\mathrm{xgb}}(f) \geq \inf\left\{\|\boldsymbol{\beta}\|_1 : \beta^{L,U}_{\mathbf{p},\mathbf{q}} \text{ satisfy (E.8)}\right\}.$$

Since (E.8) is a system of linear equations, there clearly exist minimizers $\beta^{L,U}_{\mathbf{p},\mathbf{q}}$ of the right-hand side. Let $\hat{\beta}^{L,U}_{\mathbf{p},\mathbf{q}}$ denote one such minimizer and let $\hat{\mu}_{L,U}$ be the corresponding discrete signed measures defined via (E.7).

Choose any constant $c_0 \in \mathbb{R}$. By construction, the function $f_{c_0,\{\hat{\mu}_{L,U}\}}$ is piecewise constant as in (E.1) and satisfies (E.2). By Lemma 38, there exists a constant $b \in \mathbb{R}$ such that $f(\mathbf{x}) = b + f_{c_0,\{\hat{\mu}_{L,U}\}}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$. Defining $c = b + c_0$, we then have $f_{c,\{\hat{\mu}_{L,U}\}} \equiv f$. It follows that

$$\|\hat{\boldsymbol{\beta}}\|_1 = \sum_{0<|L|+|U|\leq s} \|\hat{\mu}_{L,U}\|_{\mathrm{TV}} \geq V^{d,s}_{\infty-\mathrm{xgb}}(f) \geq \|\hat{\boldsymbol{\beta}}\|_1,$$

so that

$$V^{d,s}_{\infty-\mathrm{xgb}}(f) = \sum_{0<|L|+|U|\leq s} \|\hat{\mu}_{L,U}\|_{\mathrm{TV}}.$$

Therefore, the minimum in the definition of $V^{d,s}_{\infty-\mathrm{xgb}}(f)$ is attained by discrete signed measures. This proves $V^{d,s}_{\infty-\mathrm{xgb}}(f) = V^{d,s}_{\mathrm{xgb}}(f)$. □

Proof of Lemma 38. Since the alternating-sum functional is linear, it suffices to show that if $f \in \mathcal{F}^{d,s}_{\mathrm{st}}$ is piecewise constant as in (E.1) and satisfies

$$\Delta^S_{\mathbf{m}}(f) = 0 \quad \text{(E.9)}$$

for all $\emptyset \neq S \subseteq [d]$ with $|S| \leq s$ and $\mathbf{m} \in \prod_{j\in S}[n_j]$, then f is a constant function.

Fix such a $f \in \mathcal{F}^{d,s}_{\mathrm{st}}$. As seen in (E.3), the expansion of $\Delta^S_{\mathbf{m}}(f)$ involves the summation over $L \subseteq S \subseteq L \cup U$ with $|L| + |U| \leq s$. When $|S| > s$, this summation is vacuous, so that $\Delta^S_{\mathbf{m}}(f) = 0$ holds automatically. Thus, (E.9) in fact holds for all nonempty $S \subseteq [d]$.

For each $j \in [d]$, define

$$w^{(j)}_0 = v^{(j)}_1 - 1 \quad \text{and} \quad w^{(j)}_{m_j} = v^{(j)}_{m_j} \text{ for } m_j \in [n_j],$$

and let $\boldsymbol{\phi}$ denote the vector of evaluations of f at $(w^{(1)}_{m_1}, \ldots, w^{(d)}_{m_d})$; that is,

$$\phi(\mathbf{m}) = f(w^{(1)}_{m_1}, \ldots, w^{(d)}_{m_d}) \quad \text{for } \mathbf{m} = (m_1, \ldots, m_d) \in \prod_{j=1}^{d}\{0, \ldots, n_j\}.$$

Because f is piecewise constant as in (E.1) and right-continuous, the vector $\boldsymbol{\phi}$ completely determines f. Moreover, for the same reason, (E.9) is equivalent to

$$\Delta^S_{\mathbf{m}}\boldsymbol{\phi} := \sum_{\boldsymbol{\delta}\in\{0,1\}^{|S|}} (-1)^{\sum_{j\in S}\delta_j} \cdot \phi\big(m_j - \delta_j, j \in S; 0, j \in S^c\big) = 0 \quad \text{for } \mathbf{m} = (m_j, j \in S) \in \prod_{j\in S}[n_j] \quad \text{(E.10)}$$

for all nonempty S. Thus, it suffices to show that if $\boldsymbol{\phi}$ satisfies (E.10), then $\boldsymbol{\phi}$ is a constant vector.

We prove this claim by induction on d. The case $d = 1$ is straightforward. When $d = 1$, (E.10) reduces to

$$\Delta_m^{\{1\}}\boldsymbol{\phi} = \phi(m) - \phi(m-1) = 0 \quad \text{for } m \in [n_1].$$

Hence, in this case, it is clear that $\boldsymbol{\phi}$ is a constant vector. Suppose the claim holds for $d-1$, and let us prove it for d. For each $m_d \in \{0, \dots, n_d\}$, let $\boldsymbol{\phi}^{(m_d)}$ denote the subvector of $\boldsymbol{\phi}$ with last index m_d; that is,

$$\phi^{(m_d)}(m_1, \dots, m_{d-1}) = \phi(m_1, \dots, m_{d-1}, m_d) \quad \text{for } (m_1, \dots, m_{d-1}) \in \prod_{j=1}^{d-1}\{0, \dots, n_j\}.$$

Clearly, $\boldsymbol{\phi}^{(0)}$ satisfies (E.10) for $d-1$. Note that for $m_d \in [n_d]$,

$$\Delta^S_{(m_1,\dots,m_{d-1})}\boldsymbol{\phi}^{(m_d)} - \Delta^S_{(m_1,\dots,m_{d-1})}\boldsymbol{\phi}^{(m_d-1)} = \Delta^{S\cup\{d\}}_{(m_1,\dots,m_{d-1},m_d)}\boldsymbol{\phi} = 0$$

for every nonempty $S \subseteq [d-1]$ and $(m_1, \dots, m_{d-1}) \in \prod_{j=1}^{d-1}[n_j]$. Thus, all $\boldsymbol{\phi}^{(m_d)}$ satisfy (E.10) for $d-1$. By the induction hypothesis, it follows that each $\boldsymbol{\phi}^{(m_d)}$ is a constant vector. Lastly, taking $S = \{d\}$ in (E.10) gives

$$\Delta^{\{d\}}_{m_d}\boldsymbol{\phi} = \phi(0, \dots, 0, m_d) - \phi(0, \dots, 0, m_d - 1) = 0 \quad \text{for } m_d \in [n_d].$$

Thus, the constants in $\boldsymbol{\phi}^{(m_d)}$ are the same for all m_d, which means that $\boldsymbol{\phi}$ is a constant vector. □

E.4 Proof of Theorem 16

Since the latter part of the theorem is a direct consequence of Lemma 5, we only prove the former part concerning existence here. The proof of Theorem 15 is entirely analogous.

Proof of Theorem 16. When the signed measures $\nu_{L,U}$ satisfy condition (a) of Lemma 5, the corresponding function $f_{c,\{\nu_{L,U}\}}$ can be written as

$$f_{c,\{\nu_{L,U}\}}(x_1, \dots, x_d) = c + \sum_{(L,U,\mathbf{p},\mathbf{q})\in J} \beta^{L,U}_{\mathbf{p},\mathbf{q}} \cdot \prod_{j\in L} \mathbf{1}\big(x_j \geq (v^{(j)}_{p_j} + v^{(j)}_{p_j+1})/2\big) \cdot \prod_{j\in U} \mathbf{1}\big(x_j < (v^{(j)}_{q_j} + v^{(j)}_{q_j+1})/2\big)$$

where

$$J = \left\{(L, U, \mathbf{p}, \mathbf{q}) : L, U \subseteq [d], 0 < |L|+|U| \leq s, \mathbf{p} \in \prod_{j\in L}[n_j - 1], \text{ and } \mathbf{q} \in \prod_{j\in U}[n_j - 1]\right\} \quad \text{(E.11)}$$

and

$$\beta^{L,U}_{\mathbf{p},\mathbf{q}} = \nu_{L,U}\Big(\big\{\big((v^{(j)}_{p_j} + v^{(j)}_{p_j+1})/2, j \in L; (v^{(j)}_{q_j} + v^{(j)}_{q_j+1})/2, j \in U\big)\big\}\Big)$$

for each $(L, U, \mathbf{p}, \mathbf{q}) \in J$, with $\mathbf{p} = (p_j, j \in L)$ and $\mathbf{q} = (q_j, j \in U)$.

Let $(\hat{c}, (\hat{\beta}^{L,U}_{\mathbf{p},\mathbf{q}}, (L, U, \mathbf{p}, \mathbf{q}) \in J))$ be a solution to the optimization problem

$$\begin{aligned} \operatorname{argmin} \sum_{i=1}^{n} \Bigg(y_i - c - \sum_{(L,U,\mathbf{p},\mathbf{q}) \in J} \beta^{L,U}_{\mathbf{p},\mathbf{q}} \cdot \prod_{j \in L} \mathbf{1}\big(x_j^{(i)} \geq (v^{(j)}_{p_j} + v^{(j)}_{p_j+1})/2\big) \\ \cdot \prod_{j \in U} \mathbf{1}\big(x_j^{(i)} < (v^{(j)}_{q_j} + v^{(j)}_{q_j+1})/2\big) \Bigg)^2 \\ \text{s.t.} \sum_{(L,U,\mathbf{p},\mathbf{q}) \in J} |\beta^{L,U}_{\mathbf{p},\mathbf{q}}| \leq V. \end{aligned}$$

The existence of such a solution is immediate. Define $\hat{f}^{d,s}_{n,V} : \mathbb{R}^d \to \mathbb{R}$ by

$$\hat{f}^{d,s}_{n,V}(x_1, \dots, x_d) = \hat{c} + \sum_{(L,U,\mathbf{p},\mathbf{q}) \in J} \hat{\beta}^{L,U}_{\mathbf{p},\mathbf{q}} \cdot \prod_{j \in L} \mathbf{1}\big(x_j \geq (v^{(j)}_{p_j} + v^{(j)}_{p_j+1})/2\big) \cdot \prod_{j \in U} \mathbf{1}\big(x_j < (v^{(j)}_{q_j} + v^{(j)}_{q_j+1})/2\big).$$

By construction, $\hat{f}^{d,s}_{n,V} \in \mathcal{F}^{d,s}_{\text{st,mid}}$, and it is a solution to the problem (5.9). Moreover, Lemma 5 implies that it is also a solution to (5.6). $\square$

E.5 Proof of Lemma 5

We use the following lemma for the proof. This lemma is proved right after the proof of Lemma 5.

Lemma 39. *For every $f_{c,\{\nu_{L,U}\}} \in \mathcal{F}^{d,s}_{\infty-st}$, there exists $f_{b,\{\mu_{L,U}\}} \in \mathcal{F}^{d,s}_{st}$ with discrete signed measures $\mu_{L,U}$ supported on the lattices* (5.10) *such that*

- $f_{b,\{\mu_{L,U}\}}(\mathbf{x}^{(i)}) = f_{c,\{\nu_{L,U}\}}(\mathbf{x}^{(i)})$ *for* $i = 1, \dots, n$;
- $\sum_{0<|L|+|U|\leq s} \|\mu_{L,U}\|_{TV} \leq \sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{TV}$.

Proof of Lemma 5. For $z_1, \dots, z_n \in \mathbb{R}$, define

$$V^{d,s}_{\infty-\text{xgb}}(z_1, \dots, z_n) = \inf \Bigg\{ \sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{\text{TV}} : f_{c,\{\nu_{L,U}\}}(\mathbf{x}^{(i)}) = z_i \text{ for } i = 1, \dots, n \Bigg\}.$$

For simplicity, we suppress the dependence on the design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. By definition,

$$V^{d,s}_{\infty-\text{xgb}}(f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)})) \leq V^{d,s}_{\infty-\text{xgb}}(f) \quad \text{for every } f \in \mathcal{F}^{d,s}_{\infty-\text{st}}.$$

By Lemma 39, for each $z_1, \dots, z_n$, we have

$$V^{d,s}_{\infty-\text{xgb}}(z_1, \dots, z_n) = \inf \Bigg\{ \sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{\text{TV}} : f_{c,\{\nu_{L,U}\}}(\mathbf{x}^{(i)}) = z_i \text{ for } i = 1, \dots, n,$$

$$\text{with } \nu_{L,U} \text{ supported on the lattices (5.10)}\Big\}.$$

Recall the index set J from (E.11), and let $\mathbf{X}$ denote the $n \times |J|$ matrix with entries

$$X_{i,(L,U,\mathbf{p},\mathbf{q})} = \prod_{j \in L} \mathbf{1}\big(x_j^{(i)} \geq (v_{p_j}^{(j)} + v_{p_j+1}^{(j)})/2\big) \cdot \prod_{j \in U} \mathbf{1}\big(x_j^{(i)} < (v_{q_j}^{(j)} + v_{q_j+1}^{(j)})/2\big)$$

for $i = 1, \dots, n$ and $(L, U, \mathbf{p}, \mathbf{q}) \in J$. We parametrize discrete signed measures $\nu_{L,U}$ supported on the lattices (5.10) by

$$\nu_{L,U}\big(\{((v_{p_j}^{(j)} + v_{p_j+1}^{(j)})/2, j \in L) \times ((v_{q_j}^{(j)} + v_{q_j+1}^{(j)})/2, j \in U)\}\big) = \beta_{\mathbf{p},\mathbf{q}}^{L,U} \tag{E.12}$$

for $L, U \subseteq [d]$ with $0 < |L| + |U| \leq s$, $\mathbf{p} = (p_j, j \in L) \in \prod_{j \in L}[n_j - 1]$, and $\mathbf{q} = (q_j, j \in U) \in \prod_{j \in U}[n_j - 1]$. With this parametrization, we can express $V_{\infty-\mathrm{xgb}}^{d,s}(z_1, \dots, z_n)$ as

$$V_{\infty-\mathrm{xgb}}^{d,s}(z_1, \dots, z_n) = \inf\big\{\|\boldsymbol{\beta}\|_1 : \mathbf{X}\boldsymbol{\beta} = \mathbf{z} - c\mathbf{1} \text{ for some } c \in \mathbb{R}\big\}, \tag{E.13}$$

where $\mathbf{z} = (z_1, \dots, z_n)$, and $\mathbf{1}$ is the all-ones vector.

Next, we show that if the set

$$\mathcal{P}_{\mathbf{z}} := \big\{\boldsymbol{\beta} \in \mathbb{R}^{|J|} : \mathbf{X}\boldsymbol{\beta} = \mathbf{z} - c\mathbf{1} \text{ for some } c \in \mathbb{R}\big\} \tag{E.14}$$

is nonempty, which is clearly the case when $\mathbf{z} = (f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)}))$ for some $f \in \mathcal{F}_{\infty-\mathrm{st}}^{d,s}$, then there exists $\hat{\boldsymbol{\beta}}$ that achieves the minimum in (E.13). Fix $\mathbf{z} \in \mathbb{R}^n$ such that $\mathcal{P}_{\mathbf{z}}$ is nonempty, and suppose $\mathbf{X}\boldsymbol{\beta}_0 = \mathbf{z} - c_0\mathbf{1}$ for some $\boldsymbol{\beta}_0$ and c_0. Clearly, the infimum on the right-hand side of (E.13) remains unchanged if we further constrain $\boldsymbol{\beta}$ to satisfy $\|\boldsymbol{\beta}\|_1 \leq \|\boldsymbol{\beta}_0\|_1$:

$$V_{\infty-\mathrm{xgb}}^{d,s}(z_1, \dots, z_n) = \inf\big\{\|\boldsymbol{\beta}\|_1 : \mathbf{X}\boldsymbol{\beta} = \mathbf{z} - c\mathbf{1} \text{ for some } c \in \mathbb{R} \text{ and } \|\boldsymbol{\beta}\|_1 \leq \|\boldsymbol{\beta}_0\|_1\big\}.$$

It is also straightforward to verify that the set

$$\mathcal{P}_{\mathbf{z}} \cap \{\boldsymbol{\beta} \in \mathbb{R}^{|J|} : \|\boldsymbol{\beta}\|_1 \leq \|\boldsymbol{\beta}_0\|_1\}$$

is nonempty, closed, and bounded. Since the map $\boldsymbol{\beta} \mapsto \|\boldsymbol{\beta}\|_1$ is continuous, it follows that there exists $\hat{\boldsymbol{\beta}}$ that attains the minimum in (E.13).

Using the results established above, we now prove the lemma. Fix $f \in \mathcal{F}_{\infty-\mathrm{st}}^{d,s}$, and let $\hat{\boldsymbol{\beta}}$ be the minimizer of (E.13) for $\mathbf{z} = (f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)}))$. Let $\hat{c}$ be the corresponding constant from (E.14), and let $\nu_{L,U}$ denote the signed measures associated with $\hat{\boldsymbol{\beta}}$ via (E.12). By construction, $f_{c,\{\nu_{L,U}\}} \in \mathcal{F}_{\mathrm{st}}^{d,s}$ satisfies the first two conditions of the lemma. Moreover,

$$\begin{aligned} V_{\infty-\mathrm{xgb}}^{d,s}(f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)})) &= \|\hat{\boldsymbol{\beta}}\|_1 = \sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{\mathrm{TV}} \\ &\geq V_{\infty-\mathrm{xgb}}^{d,s}(f_{c,\{\nu_{L,U}\}}) \geq V_{\infty-\mathrm{xgb}}^{d,s}(f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)})), \end{aligned}$$

and thus,

$$V^{d,s}_{\infty-\text{xgb}}(f_{c,\{\nu_{L,U}\}}) = \sum_{0<|L|+|U|\le s} \|\nu_{L,U}\|_{\text{TV}} = V^{d,s}_{\infty-\text{xgb}}(f(\mathbf{x}^{(1)}),\dots,f(\mathbf{x}^{(n)})) \le V^{d,s}_{\infty-\text{xgb}}(f).$$

Hence, $f_{c,\{\nu_{L,U}\}} \in \mathcal{F}^{d,s}_{\text{st,mid}}$ is a function that satisfies all the desired properties. □

Proof of Lemma 39. For each $j \in [d]$, define

$$I^{(j)}_{m_j} = (v^{(j)}_{m_j}, v^{(j)}_{m_j+1}] \qquad \text{for } m_j \in [n_j - 1],$$

and

$$\underline{I}^{(j)}_{m_j} = (v^{(j)}_1, v^{(j)}_{m_j}] \text{ and } \overline{I}^{(j)}_{m_j} = (v^{(j)}_{m_j}, v^{(j)}_{n_j}] \qquad \text{for } m_j \in [n_j].$$

Also, let

$$\underline{O}^{(j)} = (-\infty, v^{(j)}_1] \text{ and } \overline{O}^{(j)} = (v^{(j)}_{n_j}, +\infty).$$

With these notations, we define discrete signed measures $\mu_{L,U}$ and a constant b as follows. For $L, U \subseteq [d]$ with $0 < |L| + |U| \le s$, let $\mu_{L,U}$ be the discrete signed measure supported on the lattice (5.10), defined by

$$\begin{aligned}
&\mu_{L,U}\big(\{((v^{(j)}_{p_j} + v^{(j)}_{p_j+1})/2, j \in L) \times ((v^{(j)}_{q_j} + v^{(j)}_{q_j+1})/2, j \in U)\}\big) \\
&\quad = \sum_{\substack{\widetilde{L},\widetilde{U}:\widetilde{L}\supseteq L,\widetilde{U}\supseteq U \\ |\widetilde{L}|+|\widetilde{U}|\le s}} \nu_{\widetilde{L},\widetilde{U}}\Big(\prod_{j\in L} I^{(j)}_{p_j} \times \prod_{j\in\widetilde{L}\setminus L} \underline{O}^{(j)} \times \prod_{j\in U} I^{(j)}_{q_j} \times \prod_{j\in\widetilde{U}\setminus U} \overline{O}^{(j)}\Big)
\end{aligned}$$

for $(p_j, j \in L) \in \prod_{j\in L}[n_j - 1]$ and $(q_j, j \in U) \in \prod_{j\in U}[n_j - 1]$. Also, let

$$b = c + \sum_{0<|L|+|U|\le s} \nu_{L,U}\Big(\prod_{j\in L} \underline{O}^{(j)} \times \prod_{j\in U} \overline{O}^{(j)}\Big).$$

By construction, for each $(m_1, \dots, m_d) \in \prod_{j=1}^d [n_j]$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^{|L|+|U|}} \prod_{j\in L} \mathbf{1}\big(v^{(j)}_{m_j} \ge l_j\big) \cdot \prod_{j\in U} \mathbf{1}\big(v^{(j)}_{m_j} < u_j\big)\, d\mu_{L,U}(\mathbf{l},\mathbf{u}) \\
&\quad = \sum_{\mathbf{r}\in\prod_{j\in L}\{1,\dots,m_j-1\}\times\prod_{j\in U}\{m_j,\dots,n_j-1\}} \mu_{L,U}\big(\{((v^{(j)}_{r_j} + v^{(j)}_{r_j+1})/2, j \in L) \times ((v^{(j)}_{r_j} + v^{(j)}_{r_j+1})/2, j \in U)\}\big) \\
&\quad = \sum_{\substack{\widetilde{L},\widetilde{U}:\widetilde{L}\supseteq L,\widetilde{U}\supseteq U \\ |\widetilde{L}|+|\widetilde{U}|\le s}} \nu_{\widetilde{L},\widetilde{U}}\Big(\prod_{j\in L} \underline{I}^{(j)}_{m_j} \times \prod_{j\in\widetilde{L}\setminus L} \underline{O}^{(j)} \times \prod_{j\in U} \overline{I}^{(j)}_{m_j} \times \prod_{j\in\widetilde{U}\setminus U} \overline{O}^{(j)}\Big).
\end{aligned}$$

It follows that for each $(m_1, \dots, m_d) \in \prod_{j=1}^d [n_j]$,

$$f_{b,\{\mu_{L,U}\}}(v^{(1)}_{m_1}, \dots, v^{(d)}_{m_d})$$

$$
\begin{aligned}
&= b + \sum_{0<|L|+|U|\le s} \sum_{\substack{\widetilde{L},\widetilde{U}: \widetilde{L}\supseteq L, \widetilde{U}\supseteq U \\ |\widetilde{L}|+|\widetilde{U}|\le s}} \nu_{\widetilde{L},\widetilde{U}}\Big(\prod_{j\in L} \underline{I}_{m_j}^{(j)} \times \prod_{j\in \widetilde{L}\setminus L} \underline{O}^{(j)} \times \prod_{j\in U} \overline{I}_{m_j}^{(j)} \times \prod_{j\in \widetilde{U}\setminus U} \overline{O}^{(j)}\Big) \\
&= c + \sum_{0<|\widetilde{L}|+|\widetilde{U}|\le s} \sum_{L\subseteq \widetilde{L}} \sum_{U\subseteq \widetilde{U}} \nu_{\widetilde{L},\widetilde{U}}\Big(\prod_{j\in L} \underline{I}_{m_j}^{(j)} \times \prod_{j\in \widetilde{L}\setminus L} \underline{O}^{(j)} \times \prod_{j\in U} \overline{I}_{m_j}^{(j)} \times \prod_{j\in \widetilde{U}\setminus U} \overline{O}^{(j)}\Big) \\
&= c + \sum_{0<|\widetilde{L}|+|\widetilde{U}|\le s} \nu_{\widetilde{L},\widetilde{U}}\Big(\prod_{j\in \widetilde{L}} \big(\underline{I}_{m_j}^{(j)} \cup \underline{O}^{(j)}\big) \times \prod_{j\in \widetilde{U}} \big(\overline{I}_{m_j}^{(j)} \cup \overline{O}^{(j)}\big)\Big) = f_{c,\{\nu_{L,U}\}}(v_{m_1}^{(1)}, \dots, v_{m_d}^{(d)}),
\end{aligned}
$$

which implies that $f_{b,\{\mu_{L,U}\}}$ agrees with $f_{c,\{\nu_{L,U}\}}$ at all design points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$. Moreover,

$$
\begin{aligned}
\sum_{0<|L|+|U|\le s} |\mu_{L,U}|(\mathbb{R}^{|L|+|U|}) &\le \sum_{0<|L|+|U|\le s} \sum_{\mathbf{p}\in\prod_{j\in L}[n_j-1]} \sum_{\mathbf{q}\in\prod_{j\in U}[n_j-1]} \sum_{\substack{\widetilde{L},\widetilde{U}: \widetilde{L}\supseteq L, \widetilde{U}\supseteq U \\ |\widetilde{L}|+|\widetilde{U}|\le s}} \\
&\qquad |\nu_{\widetilde{L},\widetilde{U}}|\Big(\prod_{j\in L} I_{p_j}^{(j)} \times \prod_{j\in \widetilde{L}\setminus L} \underline{O}^{(j)} \times \prod_{j\in U} I_{q_j}^{(j)} \times \prod_{j\in \widetilde{U}\setminus U} \overline{O}^{(j)}\Big) \\
&= \sum_{0<|L|+|U|\le s} \sum_{\substack{\widetilde{L},\widetilde{U}: \widetilde{L}\supseteq L, \widetilde{U}\supseteq U \\ |\widetilde{L}|+|\widetilde{U}|\le s}} |\nu_{\widetilde{L},\widetilde{U}}|\Big(\prod_{j\in L} \underline{I}_{n_j}^{(j)} \times \prod_{j\in \widetilde{L}\setminus L} \underline{O}^{(j)} \times \prod_{j\in U} \overline{I}_{1}^{(j)} \times \prod_{j\in \widetilde{U}\setminus U} \overline{O}^{(j)}\Big) \\
&\le \sum_{0<|\widetilde{L}|+|\widetilde{U}|\le s} \sum_{L\subseteq \widetilde{L}} \sum_{U\subseteq \widetilde{U}} |\nu_{\widetilde{L},\widetilde{U}}|\Big(\prod_{j\in L} \underline{I}_{n_j}^{(j)} \times \prod_{j\in \widetilde{L}\setminus L} \underline{O}^{(j)} \times \prod_{j\in U} \overline{I}_{1}^{(j)} \times \prod_{j\in \widetilde{U}\setminus U} \overline{O}^{(j)}\Big) \\
&= \sum_{0<|\widetilde{L}|+|\widetilde{U}|\le s} |\nu_{\widetilde{L},\widetilde{U}}|\Big(\prod_{j\in \widetilde{L}} \big(\underline{I}_{n_j}^{(j)} \cup \underline{O}^{(j)}\big) \times \prod_{j\in \widetilde{U}} \big(\overline{I}_{1}^{(j)} \cup \overline{O}^{(j)}\big)\Big) \le \sum_{0<|\widetilde{L}|+|\widetilde{U}|\le s} |\nu_{\widetilde{L},\widetilde{U}}|(\mathbb{R}^{|\widetilde{L}|+|\widetilde{U}|}).
\end{aligned}
$$

This proves that $f_{b,\{\mu_{L,U}\}}$ is the desired function satisfying the conditions of the lemma. □

E.6 Proof of (5.12)

We use the following standard result from real analysis in the proof.

Theorem 33 (Theorem 3.29 of [49])**.** *Suppose $f : \mathbb{R} \to \mathbb{R}$ has finite total variation and is right-continuous. Then, there exists a unique constant $c \in \mathbb{R}$ and a unique finite signed (Borel) measure λ on $\mathbb{R}$ such that*

$$
f(x) = c + \int \mathbf{1}(x \ge t)\, d\lambda(t) \quad \text{for } x \in \mathbb{R}. \tag{E.15}
$$

Conversely, if $f : \mathbb{R} \to \mathbb{R}$ is of the form (E.15)*, then f has finite total variation, is right-continuous, and $TV(f) = \|\lambda\|_{TV}$.*

Proof of (5.12). We first show that

$$\mathcal{F}^{1,1}_{\infty-\mathrm{st}} = \big\{f : \mathrm{TV}(f) < \infty \text{ and } f \text{ is right-continuous}\big\}$$

and that

$$V^{1,1}_{\infty-\mathrm{xgb}}(f) = \mathrm{TV}(f) \quad \text{for all } f \in \mathcal{F}^{1,1}_{\infty-\mathrm{st}}.$$

Suppose $f = f_{c,\{\nu_{L,U}\}} \in \mathcal{F}^{1,1}_{\infty-\mathrm{st}}$. Since $d = s = 1$, the only admissible pairs of (L, U) with $0 < |L| + |U| \le s$ are $(\{1\}, \emptyset)$ and $(\emptyset, \{1\})$. Thus, f can be expressed as

$$f(x) = c + \int \mathbf{1}(x \ge l)\, d\nu_{\{1\},\emptyset}(l) + \int \mathbf{1}(x < u)\, d\nu_{\emptyset,\{1\}}(u)$$

for $x \in \mathbb{R}$. Define a signed measure λ by $\lambda = \nu_{\{1\},\emptyset} - \nu_{\emptyset,\{1\}}$. Then,

$$f(x) = c + \nu_{\emptyset,\{1\}}(\mathbb{R}) + \int \mathbf{1}(x \ge l)\, d\lambda(l)$$

for $x \in \mathbb{R}$, and

$$\|\lambda\|_{\mathrm{TV}} = |\lambda|(\mathbb{R}) \le |\nu_{\{1\},\emptyset}|(\mathbb{R}) + |\nu_{\emptyset,\{1\}}|(\mathbb{R}) = \|\nu_{\{1\},\emptyset}\|_{\mathrm{TV}} + \|\nu_{\emptyset,\{1\}}\|_{\mathrm{TV}}.$$

Hence, every $f \in \mathcal{F}^{1,1}_{\infty-\mathrm{st}}$ admits the simpler representation

$$f_{c,\lambda}(x) := c + \int \mathbf{1}(x \ge l)\, d\lambda(l),$$

and its complexity $V^{1,1}_{\infty-\mathrm{xgb}}(f)$ can be computed by

$$V^{1,1}_{\infty-\mathrm{xgb}}(f) = \inf\{\|\lambda\|_{\mathrm{TV}} : f_{c,\lambda} \equiv f\}.$$

By Theorem 33, the collection of such functions $f_{c,\lambda}$ is precisely the collection of all right-continuous functions with finite total variation. Moreover, the pair (c, λ) with $f_{c,\lambda} \equiv f$ is unique and satisfies $\|\lambda\|_{\mathrm{TV}} = \mathrm{TV}(f)$. Consequently,

$$\mathcal{F}^{1,1}_{\infty-\mathrm{st}} = \big\{f : \mathrm{TV}(f) < \infty \text{ and } f \text{ is right-continuous}\big\},$$

and for every $f \in \mathcal{F}^{1,1}_{\infty-\mathrm{st}}$,

$$V^{1,1}_{\infty-\mathrm{xgb}}(f) = \mathrm{TV}(f).$$

We next prove that

$$V^{1,2}_{\infty-\mathrm{xgb}}(f) = \frac{1}{2} \cdot \big(\mathrm{TV}(f) + |\Delta(f)|\big)$$

for all $f \in \mathcal{F}^{1,2}_{\infty-\mathrm{st}} (= \mathcal{F}^{1,1}_{\infty-\mathrm{st}})$. By the same argument as above, we can show that every $f \in \mathcal{F}^{1,2}_{\infty-\mathrm{st}}$ admits the representation

$$f_{c,\lambda,\mu}(x) := c + \int \mathbf{1}(x \ge l)\, d\lambda(l) + \int \mathbf{1}(l \le x < u)\, d\mu(l, u),$$

where λ and μ are finite signed measures on $\mathbb{R}$ and $\mathbb{R}^2$, respectively, and its complexity $V^{1,2}_{\infty-\mathrm{xgb}}(f)$ is given by

$$V^{1,2}_{\infty-\mathrm{xgb}}(f) = \inf\big\{\|\lambda\|_{\mathrm{TV}} + \|\mu\|_{\mathrm{TV}} : f_{c,\lambda,\mu} \equiv f\big\}.$$

First, suppose $f \equiv f_{c,\lambda,\mu} \in \mathcal{F}^{1,2}_{\infty-\mathrm{st}}$. For every $x < y$, we have

$$\begin{aligned}|f(x) - f(y)| &= \big| -\lambda((x,y]) + \mu\big((-\infty,x] \times (x,y]\big) - \mu\big((x,y] \times (y,+\infty)\big)\big| \\ &\le |\lambda|((x,y]) + |\mu|\big(\mathbb{R} \times (x,y]\big) + |\mu|\big((x,y] \times \mathbb{R}\big),\end{aligned}$$

and it thus follows that

$$\mathrm{TV}(f) \le \|\lambda\|_{\mathrm{TV}} + 2\|\mu\|_{\mathrm{TV}}.$$

Moreover, we have

$$|\Delta(f)| = \big|\lim_{x\to+\infty} f(x) - \lim_{x\to-\infty} f(x)\big| = |\lambda(\mathbb{R})| \le \|\lambda\|_{\mathrm{TV}}.$$

Combining these two inequalities, we obtain

$$\frac{1}{2}\cdot\big(\mathrm{TV}(f) + |\Delta(f)|\big) \le \|\lambda\|_{\mathrm{TV}} + \|\mu\|_{\mathrm{TV}}.$$

Taking the infimum over all λ and μ with $f_{c,\lambda,\mu} \equiv f$ gives

$$\frac{1}{2}\cdot\big(\mathrm{TV}(f) + |\Delta(f)|\big) \le V^{1,2}_{\infty-\mathrm{xgb}}(f),$$

which proves one direction of the desired identity.

We now prove the reverse inequality. Suppose $f \in \mathcal{F}^{1,2}_{\infty-\mathrm{st}} (= \mathcal{F}^{1,1}_{\infty-\mathrm{st}})$. Since f has finite total variation and is right-continuous, Theorem 33 guarantees the existence of a constant $c \in \mathbb{R}$ and a finite signed measure λ on $\mathbb{R}$ such that

$$f(x) = c + \int \mathbf{1}(x \ge l)\, d\lambda(l) \quad \text{for } x \in \mathbb{R}$$

and $\|\lambda\|_{\mathrm{TV}} = \mathrm{TV}(f)$. Let $\lambda = \lambda^+ - \lambda^-$ be the Jordan decomposition of λ, and let (P, N) be a Hahn decomposition. Without loss of generality, we assume

$$\lambda^+(\mathbb{R}) \ge \lambda^-(\mathbb{R}).$$

Then,

$$|\Delta(f)| = \big|\lim_{x\to+\infty} f(x) - \lim_{x\to-\infty} f(x)\big| = |\lambda(\mathbb{R})| = \lambda^+(\mathbb{R}) - \lambda^-(\mathbb{R}),$$

and thus,

$$\frac{1}{2}\cdot\big(\mathrm{TV}(f) + |\Delta(f)|\big) = \frac{1}{2}\cdot\big(\lambda^+(\mathbb{R}) + \lambda^-(\mathbb{R}) + \lambda^+(\mathbb{R}) - \lambda^-(\mathbb{R})\big) = \lambda^+(\mathbb{R}).$$

Define a (Borel) measure $\widetilde{\lambda}$ on $\mathbb{R}$ by

$$\widetilde{\lambda}(E) = \Big(1 - \frac{\lambda^-(\mathbb{R})}{\lambda^+(\mathbb{R})}\Big) \cdot \lambda^+(E)$$

for Borel sets $E \subseteq \mathbb{R}$. The assumption $\lambda^+(\mathbb{R}) \geq \lambda^-(\mathbb{R})$ ensures that $\widetilde{\lambda}(E)$ is nonnegative for all E. Next, define a signed (Borel) measure μ on $\mathbb{R}^2$ by

$$\mu(E_1 \times E_2) = \frac{1}{\lambda^+(\mathbb{R})} \cdot \big(\lambda^+(E_1) \cdot \lambda^-(E_2) - \lambda^-(E_1) \cdot \lambda^+(E_2)\big)$$

for Borel sets $E_1, E_2 \subseteq \mathbb{R}$. By construction,

$$\mu(E_1 \times E_2) = \begin{cases} \lambda^+(E_1) \cdot \lambda^-(E_2)/\lambda^+(\mathbb{R}) \geq 0 & \text{if } E_1 \subseteq P, E_2 \subseteq N, \\ -\lambda^-(E_1) \cdot \lambda^+(E_2)/\lambda^+(\mathbb{R}) \leq 0 & \text{if } E_1 \subseteq N, E_2 \subseteq P, \\ 0 & \text{otherwise,} \end{cases}$$

and therefore,

$$\|\mu\|_{\mathrm{TV}} = \mu(P \times N) - \mu(N \times P) = \frac{2\lambda^+(P) \cdot \lambda^-(N)}{\lambda^+(\mathbb{R})} = 2\lambda^-(\mathbb{R}).$$

Define a signed measure $\widetilde{\mu}$ on $\mathbb{R}^2$ by

$$d\widetilde{\mu}(l, u) = 1(l \leq u) \cdot d\mu(l, u).$$

Since μ is anti-symmetric, i.e., $\mu(E_1 \times E_2) = -\mu(E_2 \times E_1)$ for all Borel sets $E_1, E_2 \subseteq \mathbb{R}$,

$$\|\widetilde{\mu}\|_{\mathrm{TV}} = \frac{1}{2} \cdot \|\mu\|_{\mathrm{TV}} = \lambda^-(\mathbb{R}).$$

Hence,

$$\|\widetilde{\lambda}\|_{\mathrm{TV}} + \|\widetilde{\mu}\|_{\mathrm{TV}} = \widetilde{\lambda}(\mathbb{R}) + \|\widetilde{\mu}\|_{\mathrm{TV}} = \lambda^+(\mathbb{R}) - \lambda^-(\mathbb{R}) + \lambda^-(\mathbb{R}) = \lambda^+(\mathbb{R}).$$

Now, observe that

$$\int \mathbf{1}(x \geq l)\, d\widetilde{\lambda}(l) = \int \mathbf{1}(x \geq l)\, d\lambda^+(l) - \frac{\lambda^-(\mathbb{R})}{\lambda^+(\mathbb{R})} \cdot \lambda^+((-\infty, x])$$

and that

$$\begin{aligned} \int \mathbf{1}(l \leq x < u)\, d\widetilde{\mu}(l, u) &= \int \mathbf{1}(l \leq x < u)\, d\mu(l, u) = \mu\big((-\infty, x] \times (x, +\infty)\big) \\ &= \frac{1}{\lambda^+(\mathbb{R})} \cdot \big(\lambda^+((-\infty, x]) \cdot \lambda^-((x, +\infty)) - \lambda^-((-\infty, x]) \cdot \lambda^+((x, +\infty))\big) \end{aligned}$$

$$= \frac{\lambda^-(\mathbb{R})}{\lambda^+(\mathbb{R})} \cdot \lambda^+((-\infty, x]) - \lambda^-((-\infty, x]) = \frac{\lambda^-(\mathbb{R})}{\lambda^+(\mathbb{R})} \cdot \lambda^+((-\infty, x]) - \int \mathbf{1}(x \geq l)\, d\lambda^-(l)$$

for every $x \in \mathbb{R}$. Combining these two equations gives

$$f_{c,\widetilde{\lambda},\widetilde{\mu}}(x) = c + \int \mathbf{1}(x \geq l)\, d\widetilde{\lambda}(l) + \int \mathbf{1}(l \leq x < u)\, d\widetilde{\mu}(l, u) = c + \int \mathbf{1}(x \geq l)\, d\lambda(l) = f(x)$$

for every $x \in \mathbb{R}$. As a result,

$$V^{1,2}_{\infty-\mathrm{xgb}}(f) \leq \|\widetilde{\lambda}\|_{\mathrm{TV}} + \|\widetilde{\mu}\|_{\mathrm{TV}} = \lambda^+(\mathbb{R}) = \frac{1}{2} \cdot \big(\mathrm{TV}(f) + |\Delta(f)|\big),$$

which proves the reverse inequality. □

E.7 Proof of Proposition 12

Proof of Proposition 12. To prove (5.14), it suffices to show that

$$\mathrm{HK}_{\mathbf{a}}(f)/\min(2^s - 1, 2^d) \leq V^{d,s}_{\infty-\mathrm{xgb}}(f) \leq \mathrm{HK}_{\mathbf{a}}(f)$$

for each anchor point $\mathbf{a} \in \{-\infty, +\infty\}^d$. Here, we prove this inequality only for the case $\mathbf{a} = (-\infty, \dots, -\infty)$. The same argument applies to the other anchor point choices.

Recall that $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ is the collection of all functions $f_{c,\{\nu_{L,U}\}}$ of the form (5.4). For each $f_{c,\{\nu_{L,U}\}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$, by modifying each atom $A^{L,U}_{\mathbf{l},\mathbf{u}}$ as

$$\begin{aligned} A^{L,U}_{\mathbf{l},\mathbf{u}}(x_1, \dots, x_d) &= \prod_{j \in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j \in U} \mathbf{1}(x_j < u_j) \\ &= \prod_{j \in L \setminus U} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j \in U \setminus L} \big(1 - \mathbf{1}(x_j \geq u_j)\big) \\ &\quad \cdot \prod_{j \in L \cap U} \big(\mathbf{1}(x_j \geq l_j) - \mathbf{1}(x_j \geq u_j)\big) \cdot \prod_{j \in L \cap U} \mathbf{1}(l_j \leq u_j), \end{aligned} \tag{E.16}$$

we can represent $f_{c,\{\nu_{L,U}\}}$ as

$$f_{c,\{\nu_{L,U}\}}(x_1, \dots, x_d) = f_{b,\{\mu_S\}}(x_1, \dots, x_d) := b + \sum_{0 < |S| \leq s} \int_{\mathbb{R}^{|S|}} \prod_{j \in S} \mathbf{1}(x_j \geq t_j)\, d\mu_S(t_j, j \in S) \tag{E.17}$$

for some $b \in \mathbb{R}$ and finite signed (Borel) measures μ_S on $\mathbb{R}^{|S|}$, where the summation runs over all nonempty $S \subseteq [d]$ with $|S| \leq s$. Specifically, each μ_S is related to $\nu_{L,U}$ by

$$\mu_S\Big(\prod_{j \in S} E_j\Big) = \sum_{\substack{L,U: L \subseteq S \subseteq L \cup U \\ |L|+|U| \leq s}} (-1)^{|(U \setminus L) \cap S|} \sum_{K \subseteq L \cap U} (-1)^{|(L \cap U) \setminus K|} \tag{E.18}$$

$$\cdot\, \bar{\nu}_{L,U}\Big(\prod_{j \in (L\setminus U)\cup K} E_j \times \mathbb{R}^{|(L\cap U)\setminus K|} \times \mathbb{R}^{|(U\setminus S)\cup K|} \times \prod_{\substack{j\in((L\cap U)\setminus K)\\ \cup((U\setminus L)\cap S)}} E_j \Big)$$

for Borel sets $E_j \subseteq \mathbb{R}$ for $j \in S$, where $\bar{\nu}_{L,U}$ are the signed measures on $\mathbb{R}^{|L|+|U|}$ defined by

$$d\bar{\nu}_{L,U}(\mathbf{l},\mathbf{u}) = \prod_{j\in L\cap U} \mathbf{1}(l_j \le u_j)\cdot d\nu_{L,U}(\mathbf{l},\mathbf{u}).$$

This relationship between μ_S and $\nu_{L,U}$ implies

$$\begin{aligned}
\sum_{0<|S|\le s} |\mu_S|(\mathbb{R}^{|S|}) &\le \sum_{0<|S|\le s} \sum_{\substack{L,U: L\subseteq S\subseteq L\cup U\\ |L|+|U|\le s}} \sum_{K\subseteq L\cap U} |\bar{\nu}_{L,U}|(\mathbb{R}^{|L|+|U|}) \\
&= \sum_{0<|L|+|U|\le s} \sum_{S: L\subseteq S\subseteq L\cup U, S\neq\emptyset} \sum_{K\subseteq L\cap U} |\bar{\nu}_{L,U}|(\mathbb{R}^{|L|+|U|}) \\
&= \sum_{0<|L|+|U|\le s} \big(\mathbf{1}(L\neq\emptyset)\cdot 2^{|U\setminus L|} + \mathbf{1}(L=\emptyset)\cdot(2^{|U\setminus L|}-1)\big)\cdot 2^{|L\cap U|}\cdot|\bar{\nu}_{L,U}|(\mathbb{R}^{|L|+|U|}) \\
&\le \min(2^s-1, 2^d)\cdot \sum_{0<|L|+|U|\le s} |\nu_{L,U}|(\mathbb{R}^{|L|+|U|}).
\end{aligned} \tag{E.19}$$

Define

$$V_{\mathbf{a}}(f) = \inf\Big\{ \sum_{0<|S|\le s} \|\mu_S\|_{\mathrm{TV}} : f_{b,\{\mu_S\}} \equiv f \Big\} \quad \text{for } f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}. \tag{E.20}$$

By (E.19), we have

$$V^{d,s}_{\infty-\mathrm{xgb}}(f) \le V_{\mathbf{a}}(f) \le \min(2^s-1,2^d)\cdot V^{d,s}_{\infty-\mathrm{xgb}}(f) \quad \text{for every } f\in\mathcal{F}^{d,s}_{\infty-\mathrm{st}}.$$

Thus, it suffices to prove that

$$V_{\mathbf{a}}(f) = \mathrm{HK}_{\mathbf{a}}(f) \quad \text{for every } f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}.$$

Fix $f \equiv f_{b,\{\mu_S\}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$. For each nonempty $S \subseteq [d]$, we have

$$f^S_{(a_j, j\in S^c)}(x_j, j\in S) = b + \sum_{\substack{R:\emptyset\neq R\subseteq S\\ |R|\le s}} \int_{\mathbb{R}^{|R|}} \prod_{j\in R} \mathbf{1}(x_j \ge t_j)\, d\mu_R(t_j, j\in R).$$

Hence, for each nonempty $T \subseteq [d]$,

$$\begin{aligned}
&\sum_{S\subseteq T} (-1)^{|T|-|S|}\cdot f^S_{(a_j,j\in S^c)}(x_j, j\in S) \\
&\qquad = \sum_{\substack{R:\emptyset\neq R\subseteq T\\ |R|\le s}} \Big(\sum_{S: R\subseteq S\subseteq T} (-1)^{|T|-|S|}\Big) \int_{\mathbb{R}^{|R|}} \prod_{j\in R} \mathbf{1}(x_j\ge t_j)\, d\mu_R(t_j, j\in R).
\end{aligned}$$

The inner sum vanishes unless $R = T$, in which case it equals 1. Therefore, if $|T| \leq s$,

$$\sum_{S \subseteq T} (-1)^{|T|-|S|} \cdot f^{S}_{(a_j, j \in S^c)}(x_j, j \in S) = \int_{\mathbb{R}^{|T|}} \prod_{j \in T} \mathbf{1}(x_j \geq t_j)\, d\mu_T(t_j, j \in T),$$

while if $|T| > s$, the expression vanishes.

Now, fix a nonempty $T \subseteq [d]$ with $|T| \leq s$. Since

$$\mathrm{Vit}(f^{T}_{(a_j, j \in T^c)}) = \mathrm{Vit}\Big((x_j, j \in T) \mapsto \sum_{S \subseteq T} (-1)^{|T|-|S|} \cdot f^{S}_{(a_j, j \in S^c)}(x_j, j \in S)\Big),$$

we have

$$\begin{aligned}
\mathrm{Vit}(f^{T}_{(a_j, j \in T^c)}) &= \mathrm{Vit}\Big((x_j, j \in T) \mapsto \int_{\mathbb{R}^{|T|}} \prod_{j \in T} \mathbf{1}(x_j \geq t_j)\, d\mu_T(t_j, j \in T)\Big) \\
&= \sup_{u_j < v_j, j \in T} \mathrm{Vit}\Big((x_j, j \in T) \mapsto \int_{\mathbb{R}^{|T|}} \prod_{j \in T} \mathbf{1}(x_j \geq t_j)\, d\mu_T(t_j, j \in T); \prod_{j \in T} [u_j, v_j]\Big) \\
&= \sup_{u_j < v_j, j \in T} \mathrm{Vit}\Big((x_j, j \in T) \mapsto \mu_T\Big(\prod_{j \in T} (u_j, x_j]\Big); \prod_{j \in T} [u_j, v_j]\Big).
\end{aligned}$$

Moreover, by Theorem 1,

$$\begin{aligned}
&\mathrm{Vit}\Big((x_j, j \in T) \mapsto \mu_T\Big(\prod_{j \in T} (u_j, x_j]\Big); \prod_{j \in T} [u_j, v_j]\Big) \\
&\quad = \mathrm{HK}_{(u_j, j \in T)}\Big((x_j, j \in T) \mapsto \mu_T\Big(\prod_{j \in T} (u_j, x_j]\Big); \prod_{j \in T} [u_j, v_j]\Big) = |\mu_T|\Big(\prod_{j \in T} (u_j, v_j]\Big).
\end{aligned}$$

Here, the first equality holds because the map vanishes on every section containing the anchor point $(u_j, j \in T)$; that is, it becomes zero whenever $x_j = u_j$ for some $j \in T$. Consequently,

$$\mathrm{Vit}(f^{T}_{(a_j, j \in T^c)}) = |\mu_T|(\mathbb{R}^{|T|}) = \|\mu_T\|_{\mathrm{TV}}.$$

Since

$$\mathrm{Vit}(f^{T}_{(a_j, j \in T^c)}) = 0$$

for all $|T| > s$, it follows that

$$\mathrm{HK}_{\mathbf{a}}(f) = \sum_{0 < |T| \leq d} \mathrm{Vit}(f^{T}_{(a_j, j \in T^c)}) = \sum_{0 < |T| \leq s} \mathrm{Vit}(f^{T}_{(a_j, j \in T^c)}) = \sum_{0 < |T| \leq s} \|\mu_T\|_{\mathrm{TV}}.$$

Thus, there is in fact no need to take the infimum in (E.20), and

$$\mathrm{HK}_{\mathbf{a}}(f) = V_{\mathbf{a}}(f),$$

which completes the proof of (5.14).

We now investigate the tightness of the inequalities in (5.14). Fix $\mathbf{a} = (-\infty, \ldots, -\infty)$. First, observe that for the function $f(x_1, \ldots, x_d) = \mathbf{1}(x_1 \geq 0)$, we have

$$V^{d,s}_{\infty-\mathrm{xgb}}(f) = 1 = \mathrm{HK}_{\mathbf{a}}(f).$$

This shows that the right inequality in (5.14) is tight.

To show that the left inequality in (5.14) is also tight, we consider two cases, depending on whether $s \leq d$ or $s > d$. In the case $s \leq d$, consider the function $f(x_1, \ldots, x_d) = \mathbf{1}(x_1, \ldots, x_s < 0)$. It is clear that $V^{d,s}_{\infty-\mathrm{xgb}}(f) = 1$. Moreover, since

$$f(x_1, \ldots, x_d) = \prod_{j=1}^{s} \big(1 - \mathbf{1}(x_j \geq 0)\big) = 1 + \sum_{l=1}^{s} (-1)^l \sum_{1 \leq j_1 < \cdots < j_l \leq s} \mathbf{1}(x_{j_1}, \ldots, x_{j_l} \geq 0),$$

it follows that

$$\mathrm{HK}_{\mathbf{a}}(f) = 2^s - 1.$$

This shows that the left inequality in (5.14) is tight when $s \leq d$. In the case $s > d$, consider the function $f(x_1, \ldots, x_d) = \mathbf{1}(-1 \leq x_1, \ldots, x_{s-d} < 0,\ x_{s-d+1}, \ldots, x_d < 0)$. Again, it is clear that $V^{d,s}_{\infty-\mathrm{xgb}}(f) = 1$. Furthermore, we can write

$$f(x_1, \ldots, x_d) = \prod_{j=1}^{s-d} \big(\mathbf{1}(x_j \geq -1) - \mathbf{1}(x_j \geq 0)\big) \cdot \prod_{j=s-d+1}^{d} \big(1 - \mathbf{1}(x_j \geq 0)\big),$$

from which we obtain

$$\mathrm{HK}_{\mathbf{a}}(f) = 2^d.$$

This shows that the left inequality in (5.14) is tight when $s > d$. □

E.8 Proof of Proposition 9

Proof of Proposition 9. First, (a) follows from the right-continuity of each atom $A^{L,U}_{\mathbf{l},\mathbf{u}}$ and the dominated convergence theorem, together with the fact that each signed measure $\nu_{L,U}$ is finite.

For (b) and (c), recall the alternative representation $f_{b,\{\mu_S\}}$ of $f_{c,\{\nu_{L,U}\}}$, given in (E.17) and introduced in the proof of Proposition 12. Since the sum in this representation ranges over all $S \subseteq [d]$ with $0 < |S| \leq s$, the function class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ enlarges as s increases. This yields (b). Since $|S|$ is always bounded by d, the class $\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ remains unchanged once $s \geq d$, which proves (c).

Lastly, (d) follows immediately from the definition. □

E.9 Proof of Proposition 11

Proof of Proposition 11. **Step 1: $f \in \mathcal{F}^{d,d}_{\infty-\mathbf{st}} \implies f$ is right-continuous, and $\mathrm{HK}_{\mathbf{a}}(f) < \infty$ for all $\mathbf{a} \in \{-\infty, +\infty\}^d$.**

This follows directly from Propositions 9 and 12.

Step 2: $f \in \mathcal{F}^{d,s}_{\infty-\mathbf{st}} \implies$ for every $\mathbf{a} \in \{-\infty, +\infty\}^d$, (5.11) holds for all subsets $S \subseteq [d]$ with $|S| > s$.

We only consider the case $\mathbf{a} = (-\infty, \cdots, -\infty)$; the argument for other choices of anchor points is entirely analogous. Suppose that $f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$. Recall from the proof of Proposition 12 that f admits the alternative representation $f \equiv f_{b,\{\mu_S\}}$ of the form (E.17) for some $b \in \mathbb{R}$ and finite signed measures μ_S on $\mathbb{R}^{|S|}$.

Fix $T \subseteq [d]$ with $|T| > s$. Using this representation, we can express $f^T_{(a_j, j \in T^c)}$ as

$$f^T_{(a_j, j\in T^c)}(x_j, j \in T) = b + \sum_{\substack{S:\emptyset \neq S \subseteq T \\ |S| \leq s}} \int_{\mathbb{R}^{|S|}} \prod_{j \in S} \mathbf{1}(x_j \geq t_j)\, d\mu_S(t_j, j \in S).$$

It follows that

$$\begin{aligned}
&\sum_{\boldsymbol{\delta} \in \{0,1\}^{|T|}} (-1)^{\sum_{j\in T}\delta_j} \cdot f^T_{(a_j, j \in T^c)}\big((1-\delta_j)w_j + \delta_j v_j, j \in T\big) \\
&\quad = \sum_{\substack{S:\emptyset \neq S \subseteq T \\ |S| \leq s}} \sum_{\boldsymbol{\delta}_S \in \{0,1\}^{|S|}} \Bigg(\sum_{\boldsymbol{\delta}_{T\setminus S} \in \{0,1\}^{|T\setminus S|}} (-1)^{\sum_{j \in T\setminus S}\delta_j} \Bigg) \\
&\qquad\qquad \cdot (-1)^{\sum_{j\in S}\delta_j} \int_{\mathbb{R}^{|S|}} \prod_{j\in S} \mathbf{1}(x_j \geq t_j)\, d\mu_S(t_j, j\in S),
\end{aligned}$$

where $\boldsymbol{\delta}_S = (\delta_j, j \in S)$ and $\boldsymbol{\delta}_{T\setminus S} = (\delta_j, j \in T \setminus S)$. In the last expression, since $|S| \leq s < |T|$, the innermost sum always vanishes. This proves that (5.11) holds for T.

Step 3: f is right-continuous, and $\mathrm{HK}_{\mathbf{a}}(f) < \infty$ for some $\mathbf{a} \in \{-\infty, +\infty\}^d \implies f \in \mathcal{F}^{d,d}_{\infty-\mathbf{st}}$.

Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is right-continuous and that $\mathrm{HK}_{\mathbf{a}}(f) < \infty$ for some $\mathbf{a} \in \{-\infty, +\infty\}^d$. Fix a nonempty $S \subseteq [d]$, and for each integer $N \geq 1$, define $g^S_N : \mathbb{R}^{|S|} \to \mathbb{R}$ by

$$g^S_N(x_j, j \in S) = \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j\in S}\delta_j} \cdot f^S_{(a_j, j\in S^c)}\big((1-\delta_j)x_j + \delta_j(-N), j \in S\big).$$

Clearly, g^S_N inherits the right-continuity of f on the coordinates $j \in S$. Moreover,

$$\mathrm{HK}_{(-N, j\in S)}\big(g^S_N; [-N,N]^{|S|}\big) = \mathrm{Vit}\big(g^S_N; [-N,N]^{|S|}\big) = \mathrm{Vit}\big(f^S_{(a_j, j\in S^c)}; [-N,N]^{|S|}\big) < \infty.$$

Here, the first equality follows from the fact that g^S_N vanishes whenever $x_j = -N$ for some $j \in S$. Hence, by Theorem 1, there exists a unique finite signed measure ν^S_N on $[-N,N]^{|S|}$

such that

$$g_N^S(x_j, j \in S) = \nu_N^S\Big(\prod_{j \in S}(-N, x_j]\Big) \quad \text{for } (x_j, j \in S) \in [-N, N]^{|S|}.$$

Here, the endpoint $-N$ could be excluded from the intervals because g_N^S becomes zero if $x_j = -N$ for some $j \in S$. Furthermore, Theorem 1 also gives

$$|\nu_N^S|([-N, N]^{|S|}) = \mathrm{Vit}\big(f^S_{(a_j, j \in S^c)}; [-N, N]^{|S|}\big) \leq \mathrm{Vit}\big(f^S_{(a_j, j \in S^c)}\big) < \infty.$$

Now, fix integers $N_2 > N_1 \geq 1$, and observe that

$$g_{N_1}^S(x_j, j \in S) = \sum_{\boldsymbol{\delta} \in \{0,1\}^{|S|}} (-1)^{\sum_{j \in S}\delta_j} \cdot g_{N_2}^S\big((1-\delta_j)x_j + \delta_j(-N_1), j \in S\big) = \nu_{N_2}^S\Big(\prod_{j \in S}(-N_1, x_j]\Big)$$

for every $(x_j, j \in S) \in [-N_1, N_1]^{|S|}$. By uniqueness of $\nu_{N_1}^S$, this means that the restriction of $\nu_{N_2}^S$ to $[-N_1, N_1]^{|S|}$ coincides with $\nu_{N_1}^S$. Hence, $\{\nu_N^S\}_{N \geq 1}$ forms a sequence of finite signed measures, where $\nu_{N_2}^S$ is an extension of $\nu_{N_1}^S$ whenever $N_2 > N_1$.

Using this sequence of signed measures, we define a finite signed (Borel) measure ν_S on $\mathbb{R}^{|S|}$ extending ν_N^S for all $N \geq 1$. Specifically, we define ν_S by

$$\nu_S(E) = \lim_{N \to \infty} \nu_N^S(E \cap [-N, N]^{|S|}) \quad \text{for Borel sets } E \subseteq \mathbb{R}^{|S|}.$$

We first verify that $\nu_S(E)$ is well-defined for each Borel set E. For integers $M > N \geq 1$,

$$\begin{aligned} \big|\nu_M^S(E \cap [-M, M]^{|S|}) - \nu_N^S(E \cap [-N, N]^{|S|})\big| &= \big|\nu_M^S\big(E \cap ([-M, M]^{|S|} \setminus [-N, N]^{|S|})\big)\big| \\ &\leq |\nu_M^S|\big([-M, M]^{|S|} \setminus [-N, N]^{|S|}\big). \end{aligned}$$

Since

$$\sup_{N \geq 1} |\nu_N^S|([-N, N]^{|S|}) \leq \mathrm{Vit}\big(f^S_{(a_j, j \in S^c)}\big) < \infty,$$

for every $\epsilon > 0$, there exists an integer $N_0 \geq 1$ such that

$$|\nu_M^S|\big([-M, M]^{|S|} \setminus [-N, N]^{|S|}\big) < \epsilon \quad \text{for all } M > N \geq N_0.$$

Thus, $\{\nu_N^S(E \cap [-N, N]^{|S|})\}_{N \geq 1}$ forms a Cauchy sequence, and hence, $\nu_S(E)$ is well-defined.

Next, we show that ν_S is countably additive. Suppose $E = \cup_{k \geq 1} E_k$ for disjoint Borel sets E_k. For integers $N_2 > N_1 \geq 1$, since $\nu_{N_2}^S$ extends $\nu_{N_1}^S$, the restriction of $|\nu_{N_2}^S|$ (the variation of $\nu_{N_2}^S$) to $[-N_1, N_1]^{|S|}$ also coincides with $|\nu_{N_1}^S|$. Thus, for each $k \geq 1$,

$$\big\{|\nu_N^S|(E_k \cap [-N, N]^{|S|})\big\}_{N \geq 1}$$

is an increasing sequence of nonnegative numbers. By the monotone convergence theorem, we have

$$\sum_{k \geq 1} \lim_{N \to \infty} |\nu_N^S|(E_k \cap [-N, N]^{|S|}) = \lim_{N \to \infty} \sum_{k \geq 1} |\nu_N^S|(E_k \cap [-N, N]^{|S|})$$

$$= \lim_{N\to\infty} |\nu_N^S|(E \cap [-N,N]^{|S|}) \leq \sup_{N\geq 1} |\nu_N^S|([-N,N]^{|S|}) \leq \mathrm{Vit}\big(f^S_{(a_j, j\in S^c)}\big) < \infty.$$

Moreover, since

$$\big|\nu_N^S\big(E_k \cap [-N,N]^{|S|}\big)\big| \leq |\nu_N^S|\big(E_k \cap [-N,N]^{|S|}\big) \leq \lim_{N\to\infty} |\nu_N^S|\big(E_k \cap [-N,N]^{|S|}\big)$$

for each k and N, the dominated convergence theorem yields

$$\begin{aligned}\nu_S(E) &= \lim_{N\to\infty} \nu_N^S(E\cap[-N,N]^{|S|}) = \lim_{N\to\infty} \sum_{k\geq 1} \nu_N^S(E_k \cap [-N,N]^{|S|}) \\ &= \sum_{k\geq 1} \lim_{N\to\infty} \nu_N^S(E_k \cap [-N,N]^{|S|}) = \sum_{k\geq 1} \nu_S(E_k).\end{aligned}$$

This establishes that ν_S is countably additive.

For each $N \geq 1$, it is clear from the definition of ν_S that for any Borel set $E \subseteq [-N,N]^{|S|}$, we have $\nu_S(E) = \nu_N^S(E)$. Furthermore,

$$|\nu_S|(\mathbb{R}^{|S|}) = \lim_{N\to\infty} |\nu_S|([-N,N]^{|S|}) = \lim_{N\to\infty} |\nu_N^S|([-N,N]^{|S|}) \leq \mathrm{Vit}\big(f^S_{(a_j, j\in S^c)}\big) < \infty.$$

Hence, ν_S is a finite signed measure on $\mathbb{R}^{|S|}$ extending ν_N^S for all $N \geq 1$, as desired.

Define $N_{\mathbf{a}} = \{j \in [d] : a_j = -\infty\}$. For each integer $N \geq 1$, we have

$$\begin{aligned}&\sum_{\boldsymbol{\delta}\in\{0,1\}^{|S|}} (-1)^{\sum_{j\in S}\delta_j} \cdot f^S_{(a_j, j\in S^c)}\big((1-\delta_j)x_j + \delta_j(-N), j \in S\cap N_{\mathbf{a}}; (1-\delta_j)x_j + \delta_j N, j \in S\setminus N_{\mathbf{a}}\big) \\ &= \sum_{\boldsymbol{\delta}\in\{0,1\}^{|S|}} (-1)^{\sum_{j\in S}\delta_j} \cdot g_N^S\big((1-\delta_j)x_j + \delta_j(-N), j \in S\cap N_{\mathbf{a}}; (1-\delta_j)x_j + \delta_j N, j \in S\setminus N_{\mathbf{a}}\big) \\ &= (-1)^{|S\setminus N_{\mathbf{a}}|} \cdot \nu_S\Big(\prod_{j\in S\cap N_{\mathbf{a}}} (-N, x_j] \times \prod_{j\in S\setminus N_{\mathbf{a}}} (x_j, N]\Big).\end{aligned}$$

Taking the limit as $N \to \infty$ yields

$$\sum_{R\subseteq S} (-1)^{|S|-|R|} \cdot f^R_{(a_j, j\in R^c)}(x_j, j\in R) = \widetilde{\nu}_S\Big(\prod_{j\in S\cap N_{\mathbf{a}}} (-\infty, x_j] \times \prod_{j\in S\setminus N_{\mathbf{a}}} (x_j, +\infty)\Big)$$

where $\widetilde{\nu}_S$ is the signed measure on $\mathbb{R}^{|S|}$ defined by $\widetilde{\nu}_S = (-1)^{|S\setminus N_{\mathbf{a}}|} \cdot \nu_S$. Moreover, since

$$f(x_1,\ldots,x_d) = \lim_{\mathbf{z}\to\mathbf{a}} f(\mathbf{z}) + \sum_{S:\emptyset\neq S\subseteq[d]} \sum_{R\subseteq S} (-1)^{|S|-|R|} \cdot f^R_{(a_j, j\in R^c)}(x_j, j\in R) \quad \text{for } (x_1,\ldots,x_d)\in\mathbb{R}^d,$$

it follows that

$$f(x_1,\ldots,x_d) = \lim_{\mathbf{z}\to\mathbf{a}} f(\mathbf{z}) + \sum_{S:\emptyset\neq S\subseteq[d]} \widetilde{\nu}_S\Big(\prod_{j\in S\cap N_{\mathbf{a}}} (-\infty, x_j] \times \prod_{j\in S\setminus N_{\mathbf{a}}} (x_j, +\infty)\Big)$$

$$= \lim_{\mathbf{z}\to\mathbf{a}} f(\mathbf{z}) + \sum_{S:\emptyset\neq S\subseteq[d]} \int_{\mathbb{R}^{|S|}} \Big(\prod_{j\in S\cap N_{\mathbf{a}}} \mathbf{1}(x_j \geq t_j)\Big) \cdot \Big(\prod_{j\in S\setminus N_{\mathbf{a}}} \mathbf{1}(x_j < t_j)\Big)\, d\widetilde{\nu}_S(t_j, j\in S)$$

for all $(x_1,\dots,x_d)\in\mathbb{R}^d$. This proves that $f\in\mathcal{F}^{d,d}_{\infty-\mathrm{st}}$.

Step 4: $f \in \mathcal{F}^{d,d}_{\infty-\mathbf{st}}$ **and for some** $\mathbf{a}\in\{-\infty,+\infty\}^d$**,** (5.11) **holds for all subsets** $S\subseteq[d]$ **with** $|S|>s \implies f\in\mathcal{F}^{d,s}_{\infty-\mathbf{st}}$**.**

Now, we assume that f additionally satisfies condition (5.11) for all $S\subseteq[d]$ with $|S|>s$ for some $\mathbf{a}\in\{-\infty,+\infty\}^d$. Since the additional condition is vacuous when $s\geq d$, we assume that $s<d$. Here, we present the argument only for the case $\mathbf{a}=(-\infty,\cdots,-\infty)$, but the proof for other anchor points is entirely analogous.

Since $f\in\mathcal{F}^{d,d}_{\infty-\mathrm{st}}$, f admits the alternative representation (E.17) for some $b\in\mathbb{R}$ and finite signed measures μ_S on $\mathbb{R}^{|S|}$. For each $S\subseteq[d]$ with $|S|>s$, we have

$$\sum_{\boldsymbol{\delta}\in\{0,1\}^{|S|}} (-1)^{\sum_{j\in S}\delta_j}\cdot f^S_{(a_j, j\in S^c)}\big((1-\delta_j)v_j+\delta_j u_j, j\in S\big) = \mu_S\Big(\prod_{j\in S}(u_j,v_j]\Big)$$

for all $u_j<v_j, j\in S$. Therefore, condition (5.11) implies that for all such S and all $u_j<v_j, j\in S$,

$$\mu_S\Big(\prod_{j\in S}(u_j,v_j]\Big)=0.$$

By Dynkin's π-λ theorem, this forces $\mu_S=0$. Hence, all integrals over μ_S with $|S|>s$ can be dropped from (E.17), and we can conclude that $f\in\mathcal{F}^{d,s}_{\infty-\mathrm{st}}$. □

E.10 Proof of Proposition 13

Proof of Proposition 13. Fix $j_0\in[d]$ and $t_{j_0}\in\mathbb{R}$. Define $g:\mathbb{R}^d\to\mathbb{R}$ as in the statement of the proposition with $j=j_0$. By symmetry, it suffices to show that $V^{d,s}_{\infty-\mathrm{xgb}}(g)\leq V^{d,s}_{\infty-\mathrm{xgb}}(f)$.

Suppose that $f\equiv f_{c,\{\nu_{L,U}\}}$. For each $L,U\subseteq[d]$ with $0<|L|+|U|\leq s$, we define a signed measure $\mu_{L,U}$ on $\mathbb{R}^{|L|+|U|}$ as follows. If $j_0\notin L\cup U$, set $\mu_{L,U}=\nu_{L,U}$. If $j_0\in L\setminus U$, define $\mu_{L,U}$ as the pushforward of $\nu_{L\setminus\{j_0\},U\cup\{j_0\}}$ under the map

$$\big((l_j, j\in L\setminus\{j_0\}),(u_j, j\in U\cup\{j_0\})\big)\mapsto\big((l_j, j\in L\setminus\{j_0\}; t_{j_0}-u_{j_0}),(u_j, j\in U)\big),$$

if $j_0\in U\setminus L$, define $\mu_{L,U}$ as the pushforward of $\nu_{L\cup\{j_0\},U\setminus\{j_0\}}$ under the map

$$\big((l_j, j\in L\cup\{j_0\}),(u_j, j\in U\setminus\{j_0\})\big)\mapsto\big((l_j, j\in L),(u_j, j\in U\setminus\{j_0\}; t_{j_0}-l_{j_0})\big),$$

and if $j_0\in L\cap U$, define $\mu_{L,U}$ as the pushforward of $\nu_{L,U}$ under the map

$$\big((l_j, j\in L),(u_j, j\in U)\big)\mapsto\big((l_j, j\in L\setminus\{j_0\}; t_{j_0}-u_{j_0}),(u_j, j\in U\setminus\{j_0\}; t_{j_0}-l_{j_0})\big).$$

With these definitions, one readily checks that $f_{c,\{\mu_{L,U}\}} \equiv g$. Moreover, there is a one-to-one correspondence between the signed measures $\mu_{L,U}$ and the signed measures $\nu_{L,U}$, under which the corresponding signed measures have the same total variation. Therefore,

$$
\begin{aligned}
V^{d,s}_{\infty-\mathrm{xgb}}(g) &\leq \sum_{0<|L|+|U|\leq s} \|\mu_{L,U}\|_{\mathrm{TV}} \\
&= \sum_{\substack{0<|L|+|U|\leq s \\ j_0 \notin L\cup U}} \|\mu_{L,U}\|_{\mathrm{TV}} + \sum_{\substack{0<|L|+|U|\leq s \\ j_0 \in L\setminus U}} \|\mu_{L,U}\|_{\mathrm{TV}} + \sum_{\substack{0<|L|+|U|\leq s \\ j_0 \in U\setminus L}} \|\mu_{L,U}\|_{\mathrm{TV}} + \sum_{\substack{0<|L|+|U|\leq s \\ j_0 \in L\cap U}} \|\mu_{L,U}\|_{\mathrm{TV}} \\
&= \sum_{\substack{0<|L|+|U|\leq s \\ j_0 \notin L\cup U}} \|\nu_{L,U}\|_{\mathrm{TV}} + \sum_{\substack{0<|L|+|U|\leq s \\ j_0 \in L\setminus U}} \|\nu_{L\setminus\{j_0\},U\cup\{j_0\}}\|_{\mathrm{TV}} \\
&\qquad + \sum_{\substack{0<|L|+|U|\leq s \\ j_0 \in U\setminus L}} \|\nu_{L\cup\{j_0\},U\setminus\{j_0\}}\|_{\mathrm{TV}} + \sum_{\substack{0<|L|+|U|\leq s \\ j_0 \in L\cap U}} \|\nu_{L,U}\|_{\mathrm{TV}} = \sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{\mathrm{TV}}.
\end{aligned}
$$

Taking the infimum over all representations $f_{c,\{\nu_{L,U}\}}$ of f yields $V^{d,s}_{\infty-\mathrm{xgb}}(g) \leq V^{d,s}_{\infty-\mathrm{xgb}}(f)$. $\square$

E.11 Proof of Theorem 17

We will use the following result, along with Lemma 21 and Theorem 25 in Appendix C, to prove the theorem. Lemma 21 provides a moment inequality for the expected supremum of multiplier empirical processes. Lemma 40 bounds the expected supremum of empirical processes with Rademacher multipliers in terms of bracketing entropy integrals. Theorem 25 reduces the problem of controlling the expected supremum of general multiplier empirical processes to the case with Rademacher multipliers. While Lemma 21 and Theorem 25 are general results, Lemma 40 is more specific to our setting. We provide the proof of Lemma 40 in Appendix E.17.

Lemma 40. *Suppose $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)}$ are i.i.d. random variables on $\mathbb{R}^d$ with density p_0, and let $\epsilon_1, \dots, \epsilon_k$ be independent Rademacher random variables, independent of $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)}$. Let $\mathcal{F}$ be a countable collection of functions from $\mathbb{R}^d$ to $\mathbb{R}$, and suppose there exist $t, D > 0$ such that $\|f\|_{p_0,2} \leq t$ and $\|f\|_\infty \leq D$ for all $f \in \mathcal{F}$. Then,*

$$
\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \Big| \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \epsilon_i f(\mathbf{x}^{(i)}) \Big|\Big] \leq C J_{[\,]}(t, \mathcal{F}, \|\cdot\|_{p_0,2}) \cdot \Big(1 + D \cdot \frac{J_{[\,]}(t, \mathcal{F}, \|\cdot\|_{p_0,2})}{t^2\sqrt{k}}\Big),
$$

where C is a universal constant, and $J_{[\,]}(t, \mathcal{F}, \|\cdot\|_{p_0,2})$ is the bracketing entropy integral defined by

$$
J_{[\,]}(t, \mathcal{F}, \|\cdot\|_{p_0,2}) = \int_0^t \sqrt{1 + \log N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_{p_0,2})}\, d\epsilon,
$$

with $N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_{p_0,2})$ denoting the ϵ-bracketing number of $\mathcal{F}$ with respect to $\|\cdot\|_{p_0,2}$.

Remark 17. *The function classes in Lemma 21, Theorem 25, and Lemma 40 are assumed to be countable to ensure measurability of the suprema inside the expectations. For an uncountable function class $\mathcal{F}$ and a stochastic process $(\Phi(f) : f \in \mathcal{F})$ indexed by $\mathcal{F}$, the supremum $\sup_{f\in\mathcal{F}} \Phi(f)$ may not be measurable.*

In the proof of Theorem 17, to avoid such a measurability issue, we define the expected supremum of Φ over $\mathcal{F}$ as

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \Phi(f)\Big] := \sup\Big\{\mathbb{E}\Big[\sup_{f\in\mathcal{G}} \Phi(f)\Big] : \mathcal{G} \subseteq \mathcal{F} \text{ is countable}\Big\},$$

following [140]. Similarly, for any $c \in \mathbb{R}$, we define

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} \Phi(f) \geq c\Big) := \sup\Big\{\mathbb{P}\Big(\sup_{f\in\mathcal{G}} \Phi(f) > c\Big) : \mathcal{G} \subseteq \mathcal{F} \text{ is countable}\Big\}.$$

With these definitions, we can avoid measurability concerns, and the above results also extend to uncountable function classes.

Proof of Theorem 17. Let $\mathcal{F}_{\mathbf{M}}(V)$ denote the collection of all functions $f_{c,\{\nu_{L,U}\}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ of the form (5.4) satisfying the following conditions:

- $\nu_{L,U}$ are supported on $\prod_{j\in L}(-M_j/2, M_j/2] \times \prod_{j\in U}(-M_j/2, M_j/2]$;
- $\sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{\mathrm{TV}} \leq V$.

It is clear from the definition of $\hat{f}^{d,s}_{n,V}$ that

$$\hat{f}^{d,s}_{n,V} \in \mathcal{F}_{\mathbf{M}}(V) \subseteq \{f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}} : V^{d,s}_{\infty-\mathrm{xgb}}(f) \leq V\}.$$

Also, the following lemma, proved in Appendix E.18, guarantees the existence of $f_{0,\mathbf{M}} \in \mathcal{F}_{\mathbf{M}}(V)$ such that $f_{0,\mathbf{M}}(\cdot) = f_0(\cdot)$ on $\prod_{j=1}^d[-M_j/2, M_j/2]$.

Lemma 41. *For every $f \in \mathcal{F}^{d,s}_{\infty-st}$ with $V^{d,s}_{\infty-xgb}(f) < V$, there exists $f_{\mathbf{M}} \in \mathcal{F}_{\mathbf{M}}(V)$ such that $f_{\mathbf{M}}(\cdot) = f(\cdot)$ on $\prod_{j=1}^d[-M_j/2, M_j/2]$.*

For each $t > 0$, define

$$D_{\mathbf{M}}(V, t) = \{f \in \mathcal{F}_{\mathbf{M}}(V) : \|f\|_{p_0,2} \leq t\}.$$

The following lemma, proved in Appendix E.19, provides a bracketing entropy integral bound for $D_{\mathbf{M}}(V, t)$, which will play a crucial role throughout the proof.

Lemma 42. *There exists a constant $C_{B,s} > 0$, depending on B and s, such that for all $t > 0$,*

$$J_{[\,]}(t, D_{\mathbf{M}}(V, t), \|\cdot\|_{p_0,2}) \leq C_{B,s} d^{\bar{s}}(1 + \log d)^{\bar{s}-1}\bigg(t \log\Big(2 + \frac{V}{t}\Big) + V^{1/2}t^{1/2}\Big[\log\Big(2 + \frac{V}{t}\Big)\Big]^{\bar{s}-1}\bigg).$$

Now, suppose we have $t_n > 4\|f_0 - f^*\|_{p_0,2}$ such that for every $r \geq 1$,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V, rt_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] \leq r\sqrt{n} t_n^2/(V+1), \\
&\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V, rt_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] \leq r\sqrt{n} t_n^2/(V+1), \text{ and} \\
&\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V, rt_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)})\Big|\Big] \leq r\sqrt{n} t_n^2/(V+1),
\end{aligned} \tag{E.21}$$

where ϵ_i are Rademacher random variables independent of $\mathbf{x}^{(i)}$, and the expectations are taken over $\mathbf{x}^{(i)}$, ϵ_i, and ξ_i. In what follows, we will first see how these bounds on the expected suprema can be used to obtain a risk bound for $\hat{f}_{n,V}^{d,s}$. The value of t_n satisfying the above inequalities will be specified in the next step, after which more precise risk bounds will be derived. The subscript n emphasizes that t_n depends on n, while its dependence on other parameters is suppressed for notational simplicity.

We first aim to bound $\mathbb{P}(\|\hat{f}_{n,V}^{d,s} - f_0\|_{p_0,2} > t)$ for $t \geq t_n$. Fix $r \geq 1$, and for each integer $j \geq 2$, define

$$\mathcal{F}_j = \big\{f \in \mathcal{F}_{\mathbf{M}}(V) : 2^{j-2} rt_n < \|f - f_{0,\mathbf{M}}\|_{p_0,2} < 2^j rt_n\big\}.$$

By construction,

$$\mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f_0\|_{p_0,2} > rt_n\big) = \mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f_{0,\mathbf{M}}\|_{p_0,2} > rt_n\big) \leq \sum_{j=2}^\infty \mathbb{P}\big(\hat{f}_{n,V}^{d,s} \in \mathcal{F}_j\big).$$

Next, let $(M_n(f) : f \in \mathcal{F}_{\mathbf{M}}(V))$ denote the stochastic processes defined by

$$M_n(f) = \frac{2}{n} \sum_{i=1}^n \xi_i (f - f^*)(\mathbf{x}^{(i)}) - \frac{1}{n} \sum_{i=1}^n (f - f^*)^2(\mathbf{x}^{(i)})$$

and define $(M(f) : f \in \mathcal{F}_{\mathbf{M}}(V))$ by

$$M(f) = -\|f - f^*\|_{p_0,2}^2.$$

Since

$$M_n(f) = \frac{2}{n} \sum_{i=1}^n \xi_i (f - f^*)(\mathbf{x}^{(i)}) - \frac{1}{n} \sum_{i=1}^n (f - f^*)^2(\mathbf{x}^{(i)}) = -\frac{1}{n} \sum_{i=1}^n \big(y_i - f(\mathbf{x}^{(i)})\big)^2 + \frac{1}{n} \sum_{i=1}^n \xi_i^2,$$

we have

$$M_n(\hat{f}_{n,V}^{d,s}) - M_n(f_{0,\mathbf{M}}) \geq 0,$$

as $\hat{f}^{d,s}_{n,V}$ minimizes the least squares over $\mathcal{F}_{\mathbf{M}}(V)$. Moreover,

$$\begin{aligned}M_n(f) - M_n(f_{0,\mathbf{M}}) &= \frac{2}{n}\sum_{i=1}^n \xi_i (f - f_{0,\mathbf{M}})(\mathbf{x}^{(i)}) - \frac{1}{n}\sum_{i=1}^n (f - f_{0,\mathbf{M}})^2(\mathbf{x}^{(i)}) \\ &\quad - \frac{2}{n}\sum_{i=1}^n (f - f_{0,\mathbf{M}})(\mathbf{x}^{(i)}) \cdot (f_{0,\mathbf{M}} - f^*)(\mathbf{x}^{(i)})\end{aligned}$$

and

$$M(f) - M(f_{0,\mathbf{M}}) = -\|f - f_{0,\mathbf{M}}\|^2_{p_0,2} - 2\mathbb{E}_{\mathbf{x}\sim p_0}\big[(f - f_{0,\mathbf{M}})(\mathbf{x}) \cdot (f_{0,\mathbf{M}} - f^*)(\mathbf{x})\big].$$

The assumption $t_n > 4\|f_0 - f^*\|_{p_0,2} = 4\|f_{0,\mathbf{M}} - f^*\|_{p_0,2}$ implies that for every $f \in \mathcal{F}_j$,

$$\begin{aligned}-(M(f) - M(f_{0,\mathbf{M}})) &\geq \|f - f_{0,\mathbf{M}}\|_{p_0,2} \cdot \big(\|f - f_{0,\mathbf{M}}\|_{p_0,2} - 2\|f_{0,\mathbf{M}} - f^*\|_{p_0,2}\big) \\ &\geq 2^{j-2} r t_n \cdot \Big(2^{j-2} r t_n - \frac{t_n}{2}\Big) \geq 2^{2j-5} r^2 t_n^2.\end{aligned}$$

Therefore,[1]

$$\begin{aligned}&\mathbb{P}\big(\hat{f}^{d,s}_{n,V} \in \mathcal{F}_j\big) \leq \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \big(M_n(f) - M_n(f_{0,\mathbf{M}})\big) \geq 0\Big) \\ &\quad\leq \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \big((M_n(f) - M_n(f_{0,\mathbf{M}})) - (M(f) - M(f_{0,\mathbf{M}}))\big) \geq 2^{2j-5} r^2 t_n^2\Big) \\ &\quad\leq \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \Big|\frac{1}{n}\sum_{i=1}^n \xi_i (f - f_{0,\mathbf{M}})(\mathbf{x}^{(i)})\Big| \geq 2^{2j-7} r^2 t_n^2\Big) \\ &\qquad + \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \Big|\frac{1}{n}\sum_{i=1}^n (f - f_{0,\mathbf{M}})^2(\mathbf{x}^{(i)}) - \|f - f_{0,\mathbf{M}}\|^2_{p_0,2}\Big| \geq 2^{2j-7} r^2 t_n^2\Big) \\ &\qquad + \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \Big|\frac{1}{n}\sum_{i=1}^n (f - f_{0,\mathbf{M}})(\mathbf{x}^{(i)}) \cdot (f_{0,\mathbf{M}} - f^*)(\mathbf{x}^{(i)}) \\ &\qquad\qquad - \mathbb{E}_{\mathbf{x}\sim p_0}\big[(f - f_{0,\mathbf{M}})(\mathbf{x}) \cdot (f_{0,\mathbf{M}} - f^*)(\mathbf{x})\big]\Big| \geq 2^{2j-8} r^2 t_n^2\Big) \\ &\quad\leq \mathbb{P}\Big(\sup_{f\in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big| \geq 2^{2j-7} r^2 \sqrt{n} t_n^2\Big) \\ &\qquad + \mathbb{P}\Big(\sup_{f\in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|^2_{p_0,2}\big)\Big| \geq 2^{2j-7} r^2 \sqrt{n} t_n^2\Big) \qquad \text{(E.22)} \\ &\qquad + \mathbb{P}\Big(\sup_{f\in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(f(\mathbf{x}^{(i)}) \cdot (f_{0,\mathbf{M}} - f^*)(\mathbf{x}^{(i)})\end{aligned}$$

[1]Because of our definitions in Remark 17, introduced to avoid measurability issues, some additional care is required in justifying the first inequality. A more detailed argument is provided in the remark following the proof.

$$- \mathbb{E}_{\mathbf{x}\sim p_0}\big[f(\mathbf{x})\cdot(f_{0,\mathbf{M}}-f^*)(\mathbf{x})\big]\Big)\Big| \geq 2^{2j-8}r^2\sqrt{n}t_n^2\Big),$$

where the second inequality uses that $-(M(f)-M(f_{0,\mathbf{M}})) \geq 2^{2j-5}r^2t_n^2$ for all $f\in\mathcal{F}_j$, and the last inequality follows because $f-f_{0,\mathbf{M}}\in D_{\mathbf{M}}(2V,2^jrt_n)$ for all $f\in\mathcal{F}_j$.

We next bound each term on the right-hand side of (E.22). As a preliminary step, we show that there exists a constant $C>0$ such that $\|f\|_\infty \leq C(V+t)$ for every $f\in D_{\mathbf{M}}(V,t)$. Suppose $f\in D_{\mathbf{M}}(V,t)$ is of the form

$$f(x_1,\dots,x_d) = c + \sum_{0<|L|+|U|\leq s}\int_{\mathbb{R}^{|L|+|U|}}\prod_{j\in L}\mathbf{1}(x_j\geq l_j)\cdot\prod_{j\in U}\mathbf{1}(x_j<u_j)\,d\nu_{L,U}(\mathbf{l},\mathbf{u})$$

where

$$\sum_{0<|L|+|U|\leq s}\|\nu_{L,U}\|_{\mathrm{TV}} \leq V.$$

Since the sum of the total variations of the signed measures is bounded by V, the second term in the above representation of f is uniformly bounded in absolute value by V. Hence, by Cauchy inequality,

$$\begin{aligned} t^2 \geq \|f\|_{p_0,2}^2 &= \int_{\prod_{j=1}^d[-M_j/2,M_j/2]} f^2(\mathbf{x})\cdot p_0(\mathbf{x})\,d\mathbf{x} \\ &\geq \int_{\prod_{j=1}^d[-M_j/2,M_j/2]}\Big(\frac{c^2}{2}-V^2\Big)\cdot p_0(\mathbf{x})\,d\mathbf{x} = \frac{c^2}{2}-V^2. \end{aligned}$$

It follows that

$$\|f\|_\infty \leq |c|+V \leq C(V+t)$$

for some universal constant $C>0$.

We now bound the first term on the right-hand side of (E.22). By Markov's inequality,

$$\begin{aligned} &\mathbb{P}\Big(\sup_{f\in D_{\mathbf{M}}(2V,2^jrt_n)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i f(\mathbf{x}^{(i)})\Big| \geq 2^{2j-7}r^2\sqrt{n}t_n^2\Big) \\ &\quad\leq \frac{1}{2^{6j-21}r^6n^{3/2}t_n^6}\cdot\mathbb{E}\Big[\sup_{f\in D_{\mathbf{M}}(2V,2^jrt_n)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i f(\mathbf{x}^{(i)})\Big|^3\Big]. \end{aligned}$$

To bound the expectation on the right, we apply Lemma 21, which gives

$$\begin{aligned} \mathbb{E}\Big[\sup_{f\in D_{\mathbf{M}}(2V,2^jrt_n)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i f(\mathbf{x}^{(i)})\Big|^3\Big] &\leq C\Big(\mathbb{E}\Big[\sup_{f\in D_{\mathbf{M}}(2V,2^jrt_n)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i f(\mathbf{x}^{(i)})\Big|\Big]\Big)^3 \\ &\quad + C\cdot 2^{3j}r^3t_n^3\|\xi_1\|_2^3 + Cn^{-3/2}\cdot\mathbb{E}\Big[\max_i\Big(|\xi_i|^3\cdot\sup_{f\in D_{\mathbf{M}}(2V,2^jrt_n)}|f(\mathbf{x}^{(i)})|^3\Big)\Big]. \end{aligned} \tag{E.23}$$

Using (E.21), the inequality

$$\max_i |\xi_i|^3 \leq \sum_{i=1}^n |\xi_i|^3,$$

and the preliminary result

$$\|f\|_\infty \leq C(V + 2^j r t_n) \text{ for all } f \in D_{\mathbf{M}}(2V, 2^j r t_n), \tag{E.24}$$

we deduce from (E.23) that

$$\begin{aligned}&\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|^3\Big] \\ &\qquad \leq C \cdot 2^{3j} r^3 n^{3/2} t_n^6 + C \cdot 2^{3j} r^3 t_n^3 \|\xi_1\|_2^3 + C n^{-1/2} \|\xi_1\|_3^3 (V^3 + 2^{3j} r^3 t_n^3).\end{aligned}$$

Substituting this back into the Markov inequality bound, we obtain

$$\begin{aligned}&\mathbb{P}\Big(\sup_{f \in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big| \geq 2^{2j-7} r^2 \sqrt{n} t_n^2\Big) \\ &\qquad \leq \frac{C}{2^{3j} r^3} + \frac{C \|\xi_1\|_2^3}{2^{3j} r^3 n^{3/2} t_n^3} + \frac{C \|\xi_1\|_3^3 V^3}{2^{6j} r^6 n^2 t_n^6} + \frac{C \|\xi_1\|_3^3}{2^{3j} r^3 n^2 t_n^3}.\end{aligned}$$

We next bound the second term on the right-hand side of (E.22). We divide into two cases depending on whether $2^j r t_n \leq V$ or not. First, suppose $2^j r t_n \leq V$. By Markov's inequality,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{f \in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{p_0,2}^2\big)\Big| \geq 2^{2j-7} r^2 \sqrt{n} t_n^2\Big) \\ &\qquad \leq \frac{1}{2^{6j-21} r^6 n^{3/2} t_n^6} \cdot \mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{p_0,2}^2\big)\Big|^3\Big].\end{aligned} \tag{E.25}$$

Also, by the standard argument of symmetrization (see, e.g., [150, Lemma 2.3.1] and [59, Theorem 16.1]),

$$\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{p_0,2}^2\big)\Big|^3\Big] \leq 8\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)})\Big|^3\Big], \tag{E.26}$$

where ϵ_i are Rademacher random variables independent of $\mathbf{x}^{(i)}$. Applying Lemma 21, we can bound the expectation on the right-hand side as

$$\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)})\Big|^3\Big] \leq C\bigg(\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(2V, 2^j r t_n)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)})\Big|\Big]\bigg)^3$$

$$+ C\Big(\sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} \|f^2\|_{p_0,2}\Big)^3 + Cn^{-3/2}\cdot \mathbb{E}\Big[\max_i \sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} |f(\mathbf{x}^{(i)})|^6\Big].$$

The contraction principle (see, e.g., [150, Proposition A.3.2] and [91, Theorem 4.12]), together with (E.21) and (E.24), gives

$$\begin{aligned}\mathbb{E}\Big[\sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)})\Big|\Big] &\le C(V+2^j r t_n)\cdot \mathbb{E}\Big[\sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] \\ &\le C(V+2^j r t_n)\cdot 2^j r\sqrt{n} t_n^2/V.\end{aligned}$$

Moreover, we have

$$\sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} \|f^2\|_{p_0,2} \le \sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} \|f\|_\infty \cdot \sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} \|f\|_{p_0,2} \le C(V+2^j r t_n)\cdot 2^j r t_n.$$

Therefore,

$$\begin{aligned}&\mathbb{E}\Big[\sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f^2(\mathbf{x}^{(i)})\Big|^3\Big] \\ &\quad\le C(V+2^j r t_n)^3\cdot 2^{3j} r^3 n^{3/2} t_n^6/V^3 + C(V+2^j r t_n)^3\cdot 2^{3j} r^3 t_n^3 + Cn^{-3/2}(V+2^j r t_n)^6.\end{aligned}$$

Combining this with (E.25) and (E.26) yields

$$\begin{aligned}&\mathbb{P}\Big(\sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{p_0,2}^2\big)\Big| \ge 2^{2j-7} r^2\sqrt{n} t_n^2\Big) \\ &\quad\le \frac{C(V+2^j r t_n)^3}{2^{3j} r^3 V^3} + \frac{C(V+2^j r t_n)^3}{2^{3j} r^3 n^{3/2} t_n^3} + \frac{C(V+2^j r t_n)^6}{2^{6j} r^6 n^3 t_n^6} \le \frac{C}{2^{3j} r^3} + \frac{CV^3}{2^{3j} r^3 n^{3/2} t_n^3} + \frac{CV^6}{2^{6j} r^6 n^3 t_n^6}.\end{aligned}$$

Next, assume that $2^j r t_n > V$. For each $f \in D_{\mathbf{M}}(2V, 2^j r t_n)$, as seen in the preliminary step, we can decompose f as $f = c + g$ where c is a constant with $|c| \le C(V + 2^j r t_n)$ and g is a function uniformly bounded in absolute value by $2V$. Using this decomposition, we can write

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{p_0,2}^2\big) = 2c\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(g(\mathbf{x}^{(i)}) - \mathbb{E}_{X\sim p_0}[g(X)]\big) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(g^2(\mathbf{x}^{(i)}) - \|g\|_{p_0,2}^2\big).$$

It follows that

$$\begin{aligned}&\mathbb{P}\Big(\sup_{f\in D_{\mathbf{M}}(2V,2^j r t_n)} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f^2(\mathbf{x}^{(i)}) - \|f\|_{p_0,2}^2\big)\Big| \ge 2^{2j-7} r^2\sqrt{n} t_n^2\Big) \\ &\quad\le \mathbb{P}\Big(\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}} \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(g(\mathbf{x}^{(i)}) - \mathbb{E}_{X\sim p_0}[g(X)]\big)\Big| \ge \frac{2^{2j-8} r^2\sqrt{n} t_n^2}{C(V+2^j r t_n)}\Big)\end{aligned}$$

$$+\,\mathbb{P}\Big(\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(g^2(\mathbf{x}^{(i)})-\|g\|_{p_0,2}^2\big)\Big|\ge 2^{2j-8}r^2\sqrt{n}t_n^2\Big). \tag{E.27}$$

By Markov's inequality, the first term of (E.27) is bounded as

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(g(\mathbf{x}^{(i)})-\mathbb{E}_{X\sim p_0}[g(X)]\big)\Big|\ge \frac{2^{2j-8}r^2\sqrt{n}t_n^2}{C(V+2^jrt_n)}\Big)\\
&\quad\le \frac{C(V+2^jrt_n)^3}{2^{6j-24}r^6n^{3/2}t_n^6}\cdot\mathbb{E}\Big[\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(g(\mathbf{x}^{(i)})-\mathbb{E}_{X\sim p_0}[g(X)]\big)\Big|^3\Big].
\end{aligned}\tag{E.28}$$

Also, by the standard argument of symmetrization,

$$\mathbb{E}\Big[\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(g(\mathbf{x}^{(i)})-\mathbb{E}_{X\sim p_0}[g(X)]\big)\Big|^3\Big]\le 8\mathbb{E}\Big[\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i g(\mathbf{x}^{(i)})\Big|^3\Big]. \tag{E.29}$$

Using Lemma 21, the expectation on the right-hand side can be bounded as

$$\mathbb{E}\Big[\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i g(\mathbf{x}^{(i)})\Big|^3\Big]\le C\bigg(\mathbb{E}\Big[\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i g(\mathbf{x}^{(i)})\Big|\Big]\bigg)^3+CV^3. \tag{E.30}$$

Applying Lemma 40 and Lemma 42, we obtain

$$\begin{aligned}
\mathbb{E}\Big[\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i g(\mathbf{x}^{(i)})\Big|\Big]&=\mathbb{E}\Big[\sup_{\substack{g\in D_{\mathbf{M}}(2V,2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i g(\mathbf{x}^{(i)})\Big|\Big]\\
&\le CJ_{[\,]}(2V,D_{\mathbf{M}}(2V,2V),\|\cdot\|_{p_0,2})\cdot\Big(1+2V\cdot\frac{J_{[\,]}(2V,D_{\mathbf{M}}(2V,2V),\|\cdot\|_{p_0,2})}{4V^2\sqrt{n}}\Big)\\
&\le C_{B,s}a_{d,s}^2V.
\end{aligned}$$

where $a_{d,s}:=d^{\bar{s}}(1+\log d)^{\bar{s}-1}$. Combining this with (E.28), (E.29), and (E.30) yields

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(g(\mathbf{x}^{(i)})-\mathbb{E}_{X\sim p_0}[g(X)]\big)\Big|\ge \frac{2^{2j-8}r^2\sqrt{n}t_n^2}{C(V+2^jrt_n)}\Big)\\
&\quad\le \frac{C_{B,s}a_{d,s}^6V^3(V+2^jrt_n)^3}{2^{6j-24}r^6n^{3/2}t_n^6}\le\frac{C_{B,s}a_{d,s}^6V^3}{2^{3j}r^3n^{3/2}t_n^3},
\end{aligned}$$

where the last inequality follows from the assumption that $2^jrt_n>V$. By a similar argument, the second term in (E.27) is bounded by

$$\mathbb{P}\Big(\sup_{\substack{g\in\mathcal{F}_{\mathbf{M}}(2V)\\ \|g\|_\infty\le 2V}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(g^2(\mathbf{x}^{(i)})-\|g\|_{p_0,2}^2\big)\Big|\ge 2^{2j-8}r^2\sqrt{n}t_n^2\Big)\le\frac{C_{B,s}a_{d,s}^6V^6}{2^{6j}r^6n^{3/2}t_n^6}\le\frac{C_{B,s}a_{d,s}^6V^3}{2^{3j}r^3n^{3/2}t_n^3}.$$

Substituting these bounds back into (E.27) gives

$$\mathbb{P}\Big(\sup_{f\in D_{\mathbf{M}}(2V,2^j rt_n)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\big(f^2(\mathbf{x}^{(i)})-\|f\|_{p_0,2}^2\big)\Big|\geq 2^{2j-7}r^2\sqrt{n}t_n^2\Big)\leq\frac{C_{B,s}a_{d,s}^6V^3}{2^{3j}r^3n^{3/2}t_n^3}.$$

Thus, whether or not $2^j rt_n\leq V$, we have

$$\begin{aligned}&\mathbb{P}\Big(\sup_{f\in D_{\mathbf{M}}(2V,2^j rt_n)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\big(f^2(\mathbf{x}^{(i)})-\|f\|_{p_0,2}^2\big)\Big|\geq 2^{2j-7}r^2\sqrt{n}t_n^2\Big)\\&\quad\leq\frac{C}{2^{3j}r^3}+\frac{C_{B,s}a_{d,s}^6V^3}{2^{3j}r^3n^{3/2}t_n^3}+\frac{CV^6}{2^{6j}r^6n^3t_n^6}.\end{aligned}$$

The third term on the right-hand side of (E.22) can be bounded similarly to the second term in the case $2^j rt_n\leq V$. Applying Markov's inequality, the symmetrization argument, and Lemma 21 in turn yields

$$\begin{aligned}&\mathbb{P}\Big(\sup_{f\in D_{\mathbf{M}}(2V,2^j rt_n)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(f(\mathbf{x}^{(i)})\cdot(f_{0,\mathbf{M}}-f^*)(\mathbf{x}^{(i)})\\&\qquad\qquad-\mathbb{E}_{\mathbf{x}\sim p_0}\big[f(\mathbf{x})\cdot(f_{0,\mathbf{M}}-f^*)(\mathbf{x})\big]\Big)\Big|\geq 2^{2j-8}r^2\sqrt{n}t_n^2\Big)\\&\quad\leq\frac{C}{2^{3j}r^3}+\frac{C\|f_0-f^*\|_{\infty,\mathbf{M}}^3}{2^{3j}r^3n^{3/2}t_n^3}+\frac{C\|f_0-f^*\|_{\infty,\mathbf{M}}^3V^3}{2^{6j}r^6n^3t_n^6},\end{aligned}$$

where $\|\cdot\|_{\infty,\mathbf{M}}$ denotes the supremum norm over $\prod_{j=1}^d[-M_j/2,M_j/2]$

$$\|g\|_{\infty,\mathbf{M}}:=\sup_{\mathbf{x}\in\prod_{j=1}^d[-M_j/2,M_j/2]}|g(\mathbf{x})|.$$

As a result, we have

$$\begin{aligned}\mathbb{P}\big(\|\hat{f}_{n,V}^{d,s}-f_0\|_{p_0,2}>rt_n\big)&=\sum_{j=2}^\infty\mathbb{P}\big(\hat{f}_{n,V}^{d,s}\in\mathcal{F}_j\big)\\&\leq\sum_{j=2}^\infty\Big[\frac{C}{2^{3j}r^3}+\frac{C(\|\xi_1\|_2^3+\|f_0-f^*\|_{\infty,\mathbf{M}}^3)}{2^{3j}r^3n^{3/2}t_n^3}+\frac{C\|\xi_1\|_3^3V^3}{2^{6j}r^6n^2t_n^6}+\frac{C\|\xi_1\|_3^3}{2^{3j}r^3n^2t_n^3}\\&\qquad+\frac{C_{B,s}a_{d,s}^6V^3}{2^{3j}r^3n^{3/2}t_n^3}+\frac{CV^3(V^3+\|f_0-f^*\|_{\infty,\mathbf{M}}^3)}{2^{6j}r^6n^3t_n^6}\Big]\\&\leq\frac{C}{r^3}+\frac{C(\|\xi_1\|_2^3+\|f_0-f^*\|_{\infty,\mathbf{M}}^3)+C_{B,s}a_{d,s}^6V^3}{r^3n^{3/2}t_n^3}+\frac{C\|\xi_1\|_3^3V^3}{r^6n^2t_n^6}+\frac{C\|\xi_1\|_3^3}{r^3n^2t_n^3}\\&\quad+\frac{CV^3(V^3+\|f_0-f^*\|_{\infty,\mathbf{M}}^3)}{r^6n^3t_n^6}.\end{aligned}$$

Plugging in $r = t/t_n$, we obtain

$$\mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f_0\|_{p_0,2} > t\big) \le \frac{Ct_n^3}{t^3} + \frac{C(\|\xi_1\|_2^3 + \|f_0 - f^*\|_{\infty,\mathbf{M}}^3) + C_{B,s}a_{d,s}^6 V^3}{n^{3/2}t^3} + \frac{C\|\xi_1\|_3^3 V^3}{n^2 t^6} + \frac{C\|\xi_1\|_3^3}{n^2 t^3} + \frac{CV^3(V^3 + \|f_0 - f^*\|_{\infty,\mathbf{M}}^3)}{n^3 t^6},$$

which holds for all $t \ge t_n$. Thus, for every $t \ge 2t_n$, we have

$$\begin{aligned}\mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f^*\|_{p_0,2} > t\big) &\le \mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f_0\|_{p_0,2} > t - \|f_0 - f^*\|_{p_0,2}\big) \le \mathbb{P}\Big(\|\hat{f}_{n,V}^{d,s} - f_0\|_{p_0,2} > \frac{t}{2}\Big)\\ &\le \frac{Ct_n^3}{t^3} + \frac{C(\|\xi_1\|_2^3 + \|f_0 - f^*\|_{\infty,\mathbf{M}}^3) + C_{B,s}a_{d,s}^6 V^3}{n^{3/2}t^3}\\ &\quad + \frac{C\|\xi_1\|_3^3 V^3}{n^2 t^6} + \frac{C\|\xi_1\|_3^3}{n^2 t^3} + \frac{CV^3(V^3 + \|f_0 - f^*\|_{\infty,\mathbf{M}}^3)}{n^3 t^6},\end{aligned}$$

where the second inequality follows from the assumption $t_n > 4\|f_0 - f^*\|_{p_0,2}$. Since

$$\int_a^\infty 2y \cdot \mathbb{P}(Y \ge y)dy = \mathbb{E}[(Y^2 - a^2)_+],$$

where $(\cdot)_+$ denotes the positive part, it follows that

$$\begin{aligned}\mathbb{E}\big[\|\hat{f}_{n,V}^{d,s} - f^*\|_{p_0,2}^2\big] &\le 4t_n^2 + \int_{2t_n}^\infty 2t \cdot \mathbb{P}\big(\|\hat{f}_{n,V}^{d,s} - f^*\|_{p_0,2} > t\big)\, dt\\ &\le Ct_n^2 + \frac{C(\|\xi_1\|_2^3 + \|f_0 - f^*\|_{\infty,\mathbf{M}}^3) + C_{B,s}a_{d,s}^6 V^3}{n^{3/2}t_n} + \frac{C\|\xi_1\|_3^3 V^3}{n^2 t_n^4}\\ &\quad + \frac{C\|\xi_1\|_3^3}{n^2 t_n} + \frac{CV^3(V^3 + \|f_0 - f^*\|_{\infty,\mathbf{M}}^3)}{n^3 t_n^4}.\end{aligned} \tag{E.31}$$

We have just seen that once we establish the bounds (E.21) on the expected suprema with some $t_n > 4\|f_0 - f^*\|_{p_0,2}$, we can bound the risk of $\hat{f}_{n,V}^{d,s}$ in terms of t_n as in the above display. Our next goal is therefore to identify a suitable t_n satisfying (E.21). To this end, we first bound

$$\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V,t)} \Big|\frac{1}{\sqrt{k}}\sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big]$$

for each $t > 0$ and $k = 1, \ldots, n$, and then apply Theorem 25 to transfer this bound to the expected supremum with ξ_i's.

Fix $k \in \{1, \ldots, n\}$. Since $\|f\|_\infty \le C(V + t)$ for every $f \in D_{\mathbf{M}}(V,t)$, Lemma 40 gives

$$\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V,t)} \Big|\frac{1}{\sqrt{k}}\sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big]$$

$$\leq C J_{[\,]}(t, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2}) \cdot \Big(1 + C(V+t) \cdot \frac{J_{[\,]}(t, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2})}{t^2\sqrt{k}}\Big).$$

Applying the entropy integral bound from Lemma 42, we obtain

$$\begin{aligned}
\mathbb{E}\Big[\sup_{f\in D_{\mathbf{M}}(V,t)} \Big|\frac{1}{\sqrt{k}}\sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] &\leq C_{B,s} a_{d,s}\Big(t\log\Big(2+\frac{V}{t}\Big) + V^{1/2}t^{1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1}\Big) \\
&\quad \cdot \Big[1 + C_{B,s}a_{d,s}(V+t)k^{-1/2}t^{-2}\Big(t\log\Big(2+\frac{V}{t}\Big) + V^{1/2}t^{1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1}\Big)\Big] \\
&\leq C_{B,s}a_{d,s}t\log\Big(2+\frac{V}{t}\Big) + C_{B,s}a_{d,s}V^{1/2}t^{1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1} \\
&\quad + C_{B,s}a_{d,s}^2 k^{-1/2}V^{3/2}t^{-1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}} + C_{B,s}a_{d,s}^2 k^{-1/2}V^2 t^{-1}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{2(\bar{s}-1)} \\
&\quad + C_{B,s}a_{d,s}^2 k^{-1/2}t\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^2 + C_{B,s}a_{d,s}^2 k^{-1/2}V^{1/2}t^{1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}} \\
&\quad + C_{B,s}a_{d,s}^2 k^{-1/2}V\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\max(2,2(\bar{s}-1))}. \qquad \text{(E.32)}
\end{aligned}$$

Recall that $a_{d,s} = d^{\bar{s}}(1+\log d)^{\bar{s}-1}$. Let $\Psi : \mathbb{R} \to \mathbb{R}$ denote the function given by the right-hand side of (E.32). A direct calculation shows that if

$$\begin{aligned}
t \geq \max\Big(& C_{B,s}a_{d,s}(V+1)k^{-1/2}\log(2+k), C_{B,s}a_{d,s}^{2/3}(V+1)k^{-1/3}[\log(2+k)]^{2(\bar{s}-1)/3}, \\
& C_{B,s}a_{d,s}^{4/5}(V+1)k^{-2/5}[\log(2+k)]^{2\bar{s}/5}, C_{B,s}a_{d,s}^2(V+1)k^{-1}[\log(2+k)]^2, \\
& C_{B,s}a_{d,s}^{4/3}(V+1)k^{-2/3}[\log(2+k)]^{2\bar{s}/3}, C_{B,s}a_{d,s}(V+1)k^{-1/2}[\log(2+k)]^{\max(1,\bar{s}-1)}\Big),
\end{aligned}$$

then

$$\Psi(t) \leq \sqrt{k}t^2/(V+1).$$

To simplify this maximum, observe that for suitable constants $C, C_s > 0$, we have

$$\begin{aligned}
&\log(2+x) \leq x^{1/6} \quad \text{for all } x \geq C, \\
&[\log(2+x)]^{2\bar{s}/5} \leq x^{1/15} \quad \text{for all } x \geq C_s, \\
&[\log(2+x)]^2 \leq x^{2/3} \quad \text{for all } x \geq C, \\
&[\log(2+x)]^{2\bar{s}/3} \leq x^{1/3} \quad \text{for all } x \geq C_s, \text{ and} \\
&[\log(2+x)]^{\max(1,\bar{s}-1)} \leq x^{1/6} \quad \text{for all } x \geq C_s.
\end{aligned}$$

Using these inequalities, we can bound terms in the maximum as follows:

$$\begin{aligned}
&k^{-1/2}\log(2+k) \leq Ck^{-1/2} + k^{-1/2}k^{1/6} \leq Ck^{-1/3}, \\
&k^{-2/5}[\log(2+k)]^{2\bar{s}/5} \leq C_s k^{-2/5} + k^{-2/5}k^{1/15} \leq C_s k^{-1/3},
\end{aligned}$$

$$k^{-1}[\log(2+k)]^2 \le Ck^{-1} + k^{-1}k^{2/3} \le Ck^{-1/3},$$
$$k^{-2/3}[\log(2+k)]^{2\bar{s}/3} \le C_s k^{-2/3} + k^{-2/3}k^{1/3} \le C_s k^{-1/3}, \text{ and}$$
$$k^{-1/2}[\log(2+k)]^{\max(1,\bar{s}-1)} \le C_s k^{-1/2} + k^{-1/2}k^{1/6} \le C_s k^{-1/3}.$$

Thus, if we set

$$\widetilde{t}_k = C_{B,s} a_{d,s}^2 (V+1) k^{-1/3} [\log(2+n)]^{2(\bar{s}-1)/3},$$

then we have

$$\Psi(\widetilde{t}_k) \le \sqrt{k}\widetilde{t}_k^2/(V+1).$$

Since the map $t \mapsto \Psi(t)/t$ is decreasing, it follows that

$$\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V, r\widetilde{t}_k)} \Big|\frac{1}{\sqrt{k}} \sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] \le \Psi(r\widetilde{t}_k) \le r\Psi(\widetilde{t}_k) \le r\sqrt{k}\widetilde{t}_k^2/(V+1)$$

for every $r \ge 1$.

We have just shown that for each $k = 1, \dots, n$,

$$\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V, r\widetilde{t}_k)} \Big|\sum_{i=1}^k \epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] \le rk\widetilde{t}_k^2/(V+1)$$

for every $r \ge 1$. By Theorem 25, this bound transfers to the expected supremum with ξ_i, giving

$$\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V, r\widetilde{t}_n)} \Big|\sum_{i=1}^n \xi_i f(\mathbf{x}^{(i)})\Big|\Big] \le 4\|\xi_1\|_{3,1} \cdot rn\widetilde{t}_n^2/(V+1)$$

for every $r \ge 1$. Therefore, redefining $\widetilde{t}_n$ by multiplying it by the factor $(1 + 4\|\xi_1\|_{3,1})$ yields the first two inequalities in (E.21) (with $t_n = \widetilde{t}_n$) for $r \ge 1$.

We now bound

$$\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V,t)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)})\Big|\Big]$$

for each $t > 0$. By following the proof of Lemma 40 with minimal modifications, we can show that

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{f \in D_{\mathbf{M}}(V,t)} \Big|\frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(\mathbf{x}^{(i)}) \cdot (f_0 - f^*)(\mathbf{x}^{(i)})\Big|\Big] \\
&\quad \le C\|f_0 - f^*\|_{\infty,\mathbf{M}} \cdot J_{[\,]}(t, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2}) \cdot \Big(1 + C(V+t) \cdot \frac{J_{[\,]}(t, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2})}{t^2\sqrt{n}}\Big).
\end{aligned}$$

Hence, repeating the computations above, we find that if we define $\bar{t}_n$ as

$$\bar{t}_n = (1 + \|f_0 - f^*\|_{\infty,\mathbf{M}}) \cdot C_{B,s} a_{d,s}^2 (V+1) n^{-1/3} [\log(2+n)]^{2(\bar{s}-1)/3},$$

then

$$\mathbb{E}\Big[\sup_{f\in D_{\mathbf{M}}(V,\bar{t}_n)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(\mathbf{x}^{(i)})\cdot(f_0-f^*)(\mathbf{x}^{(i)})\Big|\Big]\leq\sqrt{n}\bar{t}_n^2/(V+1),$$

from which the last inequality in (E.21) (with $t_n=\bar{t}_n$) follows for all $r\geq 1$.

Using $\widetilde{t}_n$ and $\bar{t}_n$, we define t_n as

$$t_n=4\|f_0-f^*\|_{p_0,2}+\max(\widetilde{t}_n,\bar{t}_n).$$

Then, for every $r\geq 1$,

$$\mathbb{E}\Big[\sup_{f\in D_{\mathbf{M}}(V,rt_n)}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(\mathbf{x}^{(i)})\Big|\Big]\leq(rt_n/\widetilde{t}_n)\cdot\sqrt{n}\widetilde{t}_n^2/(V+1)\leq r\sqrt{n}t_n^2/(V+1)$$

and the remaining two inequalities in (E.21) follow by the same argument. Hence, t_n satisfies all inequalities in (E.21) for all $r\geq 1$, and we use this t_n to derive a risk bound for $\hat{f}_{n,V}^{d,s}$.

As a last step, we substitute our t_n into (E.31). This yields

$$\begin{aligned}\mathbb{E}\big[\|\hat{f}_{n,V}^{d,s}-f^*\|_{p_0,2}^2\big]\leq C\|f_0-f^*\|_{p_0,2}^2+a_{d,s}^4(V+1)^2\Big[&C_{B,s}(1+\|\xi_1\|_{3,1}^2+\|f_0-f^*\|_{\infty,\mathbf{M}}^2)\\&\cdot n^{-2/3}[\log(2+n)]^{4(\bar{s}-1)/3}+O(n^{-2/3})\Big],\end{aligned}$$

when $s\geq 2$. On the other hand, if $s=1$, we obtain

$$\begin{aligned}\mathbb{E}\big[\|\hat{f}_{n,V}^{d,s}-f^*\|_{p_0,2}^2\big]&\leq C\|f_0-f^*\|_{p_0,2}^2\\&\quad+a_{d,1}^4(V+1)^2\Big[C\Big(\frac{\max(\widetilde{t}_n,\bar{t}_n)}{a_{d,1}^2(V+1)}\Big)^2+\frac{C\|\xi_1\|_3^3V^3}{a_{d,1}^4(V+1)^2n^2(\max(\widetilde{t}_n,\bar{t}_n))^4}+o(n^{-2/3})\Big]\\&=C\|f_0-f^*\|_{p_0,2}^2+a_{d,1}^4(V+1)^2\cdot O(n^{-2/3}),\end{aligned}$$

where the constant factors underlying $O(\cdot)$ depend on B,s, the moments of ξ_i, and $\|f_0-f^*\|_{\infty,\mathbf{M}}$. □

Remark 18. *For the first inequality in* (E.22)*, we in fact need to show that*

$$\mathbb{P}\big(\hat{f}_{n,V}^{d,s}\in\mathcal{F}_j\big)\leq\sup\Big\{\mathbb{P}\Big(\sup_{f\in\mathcal{G}}\big(M_n(f)-M_n(f_{0,\mathbf{M}})\big)\geq 0\Big):\mathcal{G}\subseteq\mathcal{F}_j\text{ is countable}\Big\},$$

since $\mathcal{F}_j$ may not be countable. Here, we give a more careful argument for this.

For each integer $N\geq 1$, let $\mathcal{G}_N$ denote the subcollection of $\mathcal{F}_j$ consisting of all $f_{c,\{\nu_{L,U}\}}$ (of the form (5.4)*) that additionally satisfy the following two conditions:*

- *$\nu_{L,U}$ are supported on*

$$\prod_{j\in L}((1/N)\mathbb{Z}\cap(-M_j/2,M_j/2])\times\prod_{j\in U}((1/N)\mathbb{Z}\cap(-M_j/2,M_j/2]),$$

where $(1/N)\mathbb{Z}:=\{m/N:m\in\mathbb{Z}\}$;

- *$c \in \mathbb{Q}$ and*
$$\nu_{L,U}(\{(p_j, j \in L) \times (q_j, j \in U)\}) \in \mathbb{Q}$$
for every $(p_j, j \in L) \times (q_j, j \in U) \in \mathbb{R}^{|L|+|U|}$.

Clearly, each $\mathcal{G}_N$ is countable, and thus, $\mathcal{G} := \cup_{N \geq 1} \mathcal{G}_N$ is countable as well. Since $\hat{f}^{d,s}_{n,V}$ is constructed from discrete signed measures with finite support, it can be easily shown that there exists a sequence $\{g_N\}_{N \geq 1}$ with $g_N \in \mathcal{G}_N \subseteq \mathcal{G}$ such that $g_N(\mathbf{x}) \to \hat{f}^{d,s}_{n,V}(\mathbf{x})$ as $N \to \infty$ for every $\mathbf{x} \in \mathbb{R}^d$. Hence, if $\hat{f}^{d,s}_{n,V} \in \mathcal{F}_j$, then

$$\sup_{f \in \mathcal{G}} \big(M_n(f) - M_n(f_{0,\mathbf{M}})\big) \geq \lim_{N \to \infty} \big(M_n(g_N) - M_n(f_{0,\mathbf{M}})\big) = M_n(\hat{f}^{d,s}_{n,V}) - M_n(f_{0,\mathbf{M}}) \geq 0.$$

Consequently,

$$\begin{aligned}\mathbb{P}\big(\hat{f}^{d,s}_{n,V} \in \mathcal{F}_j\big) &\leq \mathbb{P}\Big(\sup_{f \in \mathcal{G}} \big(M_n(f) - M_n(f_{0,\mathbf{M}})\big) \geq 0\Big) \\ &\leq \sup\Big\{\mathbb{P}\Big(\sup_{f \in \mathcal{H}} \big(M_n(f) - M_n(f_{0,\mathbf{M}})\big) \geq 0\Big) : \mathcal{H} \subseteq \mathcal{F}_j \text{ is countable}\Big\}.\end{aligned}$$

E.12 Proof of Lemma 6

Proof of Lemma 6. For each $f_{c,\{\nu_{L,U}\}} \in D_{\mathbf{M}}(V, t)$, by modifying each atom $A^{L,U}_{\mathbf{l},\mathbf{u}}$ as in (E.16), we can express $f_{c,\{\nu_{L,U}\}}$ as

$$f_{c,\{\nu_{L,U}\}}(x_1, \ldots, x_d) = f_{b,\{\mu_S\}}(x_1, \ldots, x_d) := b + \sum_{0<|S| \leq s} \int_{\mathbb{R}^{|S|}} \prod_{j \in S} \mathbf{1}(x_j \geq l_j)\, d\mu_S(l_j, j \in S) \tag{E.33}$$

for some $b \in \mathbb{R}$ and finite signed measures μ_S on $\mathbb{R}^{|S|}$, related to the original measures $\nu_{L,U}$ through (E.18). Here, the summation runs over all nonempty subsets $S \subseteq [d]$ with $|S| \leq s$. Since each $\nu_{L,U}$ is supported on $\prod_{j \in L}(-M_j/2, M_j/2] \times \prod_{j \in U}(-M_j/2, M_j/2]$, the relation (E.18) implies that each μ_S is supported on $\prod_{j \in S}(-M_j/2, M_j/2]$. Moreover, by (E.19),

$$\sum_{0<|S| \leq s} \|\mu_S\|_{\mathrm{TV}} \leq \min(2^s - 1, 2^d) \cdot \sum_{0<|L|+|U| \leq s} \|\nu_{L,U}\|_{\mathrm{TV}} \leq (2^s - 1)V \leq C_s V.$$

Hence, if we define $\widetilde{D}_{\mathbf{M}}(V, t)$ as the collection of all functions $f_{b,\{\mu_S\}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ with $\|f_{b,\{\mu_S\}}\|_{p_0,2} \leq t$ such that μ_S are supported on $\prod_{j \in S}(-M_j/2, M_j/2]$ and satisfy

$$\sum_{0<|S| \leq s} \|\mu_S\|_{\mathrm{TV}} \leq V,$$

then we have $D_{\mathbf{M}}(V, t) \subseteq \widetilde{D}_{\mathbf{M}}(C_s V, t)$.

We now split $\widetilde{D}_{\mathbf{M}}(V,t)$ into pieces, compute the bracketing entropy of each piece, and then put them together to obtain a bracketing entropy bound for $\widetilde{D}_{\mathbf{M}}(V,t)$, which will in turn yield a bound for the bracketing entropy of $D_{\mathbf{M}}(V,t)$. For every $f_{b,\{\mu_S\}} \in \widetilde{D}_{\mathbf{M}}(V,t)$, by repeating the argument (using Cauchy inequality) in the proof of Theorem 17, we can show that $|b| \le C(V+t)$ for some constant $C>0$. Set $K = \lfloor C(V+t)/\epsilon \rfloor$, and for each $k = -(K+1),\dots,K$, let $\mathcal{G}_k$ denote the collection of all functions $f_{b,\{\mu_S\}}$ of the form (E.33) with $k\epsilon \le b \le (k+1)\epsilon$ and with signed measures μ_S supported on $\prod_{j\in S}(-M_j/2, M_j/2]$ and satisfying $\sum_{0<|S|\le s} \|\mu_S\|_{\mathrm{TV}} \le V$. It is clear that

$$\widetilde{D}_{\mathbf{M}}(V,t) \subseteq \bigcup_{k=-(K+1),\dots,K} \mathcal{G}_k,$$

and hence,

$$\begin{aligned}
\log N_{[\,]}(\epsilon, \widetilde{D}_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2}) &\le \log\Big(\sum_{k=-(K+1),\dots,K} N_{[\,]}(\epsilon, \mathcal{G}_k, \|\cdot\|_{p_0,2})\Big) \\
&\le \log\Big(2 + \frac{C(V+t)}{\epsilon}\Big) + \sup_k \log N_{[\,]}(\epsilon, \mathcal{G}_k, \|\cdot\|_{p_0,2}) \qquad \text{(E.34)} \\
&\le \log\Big(2 + \frac{C(V+t)}{\epsilon}\Big) + \log N_{[\,]}(\epsilon, \mathcal{G}_0, \|\cdot\|_{p_0,2}).
\end{aligned}$$

Now, let $\mathcal{G}_\emptyset$ denote the collection of all constant functions on $\mathbb{R}^d$ with values in $[0,\epsilon]$. Also, for each nonempty $S \subseteq [d]$ with $|S| \le s$, define $\mathcal{G}_S$ as the collection of all functions on $\mathbb{R}^d$ of the form

$$(x_1,\dots,x_d) \mapsto \int_{\prod_{j\in S}[-M_j/2, M_j/2]} \prod_{j\in S} \mathbf{1}(x_j \ge l_j)\, d\mu_S(l_j, j\in S),$$

where μ_S is a finite signed measure on $\prod_{j\in S}[-M_j/2, M_j/2]$ with $\|\mu_S\|_{\mathrm{TV}} \le V$. By construction,

$$\mathcal{G}_0 \subseteq \mathcal{G}_\emptyset + (+_{0<|S|\le s} \mathcal{G}_S),$$

where $A+B = \{a+b : a\in A, b\in B\}$. It follows that

$$\log N_{[\,]}(\epsilon, \mathcal{G}_0, \|\cdot\|_{p_0,2}) \le C_s(1+\log d) + \sum_{0<|S|\le s} \log N_{[\,]}\big(\epsilon/(C_s d^{\bar{s}}), \mathcal{G}_S, \|\cdot\|_{p_0,2}\big), \qquad \text{(E.35)}$$

where $\bar{s} = \min(s,d)$.

To proceed, for each nonempty $S \subseteq [d]$ with $|S| \le s$, let $\widetilde{\mathcal{G}}_S$ denote the collection of all functions on the truncated section $\prod_{j\in S}[-M_j/2, M_j/2]$ of the original domain $\mathbb{R}^d$, obtained by restricting to the coordinates indexed by S, of the form

$$(x_j, j\in S) \mapsto \int_{\prod_{j\in S}[-M_j/2, M_j/2]} \prod_{j\in S} \mathbf{1}(x_j \ge l_j)\, d\mu_S(l_j, j\in S),$$

where μ_S is as in the definition of $\mathcal{G}_S$. Since p_0 is uniformly bounded by $B/\prod_{j=1}^d M_j$,

$$N_{[\,]}(\epsilon, \mathcal{G}_S, \|\cdot\|_{p_0,2}) \leq N_{[\,]}\Big(\Big(\frac{\prod_{j\in S} M_j}{B}\Big)^{1/2}\epsilon, \widetilde{\mathcal{G}}_S, \|\cdot\|_2\Big). \tag{E.36}$$

Furthermore, for each nonempty $S \subseteq [d]$ with $|S| \leq s$, let $\bar{\mathcal{G}}_S$ denote the collection of all functions on the scaled domain $[0,1]^{|S|}$ of the form

$$(x_j, j \in S) \mapsto \int_{[0,1]^{|S|}} \prod_{j\in S} \mathbf{1}(x_j \geq l_j)\, d\mu_S(l_j, j \in S),$$

where μ_S is a finite signed measure on $[0,1]^{|S|}$ with $\|\mu_S\|_{\mathrm{TV}} \leq V$. Through a straightforward scaling argument, it can be readily verified that

$$N_{[\,]}(\epsilon, \widetilde{\mathcal{G}}_S, \|\cdot\|_2) \leq N_{[\,]}\Big(\Big(\frac{1}{\prod_{j\in S} M_j}\Big)^{1/2}\epsilon, \bar{\mathcal{G}}_S, \|\cdot\|_2\Big) = N_{[\,]}\Big(\Big(\frac{1}{\prod_{j\in S} M_j}\Big)^{1/2}\epsilon, \bar{\mathcal{G}}_{[|S|]}, \|\cdot\|_2\Big). \tag{E.37}$$

Next, for each integer $m \geq 1$ and $R > 0$, let $\mathcal{H}_m(R)$ denote the collection of all functions on $[0,1]^m$ of the form

$$(x_1, \dots, x_m) \mapsto \int_{[0,1]^m} \prod_{j=1}^m \mathbf{1}(x_j \geq l_j)\, d\mu(l_1, \dots, l_m)$$

where μ is a finite measure (not a signed measure) on $[0,1]^m$ with $\|\mu\|_{\mathrm{TV}} \leq R$. It was proved in [55, Theorem 1.1] that

$$\log N_{[\,]}(\epsilon, \mathcal{H}_m(R), \|\cdot\|_2) \leq C_m\Big(2 + \frac{R}{\epsilon}\Big)\Big[\log\Big(2 + \frac{R}{\epsilon}\Big)\Big]^{2(m-1)}.$$

By the Jordan decomposition of signed measures,

$$\bar{\mathcal{G}}_{[|S|]} \subseteq \mathcal{H}_{|S|}(V) - \mathcal{H}_{|S|}(V),$$

where $A - B = \{a - b : a \in A, b \in B\}$. It follows that

$$\log N_{[\,]}(\epsilon, \bar{\mathcal{G}}_{[|S|]}, \|\cdot\|_2) \leq 2\log N_{[\,]}\Big(\frac{\epsilon}{2}, \mathcal{H}_{|S|}(V), \|\cdot\|_2\Big) \leq C_{|S|}\Big(2 + \frac{2V}{\epsilon}\Big)\Big[\log\Big(2 + \frac{2V}{\epsilon}\Big)\Big]^{2(|S|-1)}.$$

Substituting this bound back into (E.37), (E.36), (E.35), and (E.34) in turn, we obtain

$$\begin{aligned}\log N_{[\,]}(\epsilon, \widetilde{D}_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2}) &\leq \log\Big(2 + \frac{C(V+t)}{\epsilon}\Big) \\ &\quad + C_{B,s} d^{2\bar{s}}(1+\log d)^{2(\bar{s}-1)}\Big(2 + \frac{V}{\epsilon}\Big)\Big[\log\Big(2 + \frac{V}{\epsilon}\Big)\Big]^{2(\bar{s}-1)}.\end{aligned}$$

Lastly, since $D_{\mathbf{M}}(V,t) \subseteq \widetilde{D}_{\mathbf{M}}(C_s V, t)$, we arrive at

$$\begin{aligned}\log N_{[\,]}(\epsilon, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2}) &\leq \log\Big(2 + \frac{C_s(V+t)}{\epsilon}\Big) \\ &\quad + C_{B,s} d^{2\bar{s}}(1+\log d)^{2(\bar{s}-1)}\Big(2 + \frac{V}{\epsilon}\Big)\Big[\log\Big(2 + \frac{V}{\epsilon}\Big)\Big]^{2(\bar{s}-1)},\end{aligned}$$

which completes the proof. □

E.13 Proof of Corollary 1

Proof of Corollary 1. Observe from the proof of Theorem 17 that the risk bound for $\hat{f}^{d,s}_{n,V}$ depends continuously on V. Consequently, the desired bound in the corollary follows directly from the risk bound for $\hat{f}^{d,s}_{n,V}$ and the following inequality:

$$\mathfrak{M}^{d,s}_{n,V} \leq \liminf_{\epsilon \to 0+} \sup_{\substack{f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}} \\ V^{d,s}_{\infty-\mathrm{xgb}}(f^*) \leq V}} \mathbb{E}\|\hat{f}^{d,s}_{n,V+\epsilon} - f^*\|^2_{p_0,2}.$$

□

E.14 Proof of Theorem 18

Theorem 18 is proved similarly to Theorem 8 in Chapter 3. As in the corresponding proof in Appendix C, our argument builds on the proof ideas of [46, Theorem 4.6], which itself is motivated by those in [13, Section 4]. Following [46] and the proof of Theorem 8, we use Assouad's lemma in the following form.

Lemma 43 (Lemma 24.3 of [149])**.** *Suppose q is a positive integer, and we have $f_{\boldsymbol{\eta}} \in \mathcal{F}^{d,s}_{\infty-st}$ with $V^{d,s}_{\infty-xgb}(f_{\boldsymbol{\eta}}) \leq V$ for each $\boldsymbol{\eta} \in \{-1,1\}^q$. Then, we have the following lower bound for the minimax risk $\mathfrak{M}^{d,s}_{n,V}$:*

$$\mathfrak{M}^{d,s}_{n,V} \geq \frac{q}{8} \cdot \min_{\boldsymbol{\eta} \neq \boldsymbol{\eta}'} \frac{\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|^2_{p_0,2}}{H(\boldsymbol{\eta}, \boldsymbol{\eta}')} \cdot \min_{H(\boldsymbol{\eta},\boldsymbol{\eta}')=1} \left(1 - \sqrt{\frac{1}{2}\mathbb{E}\big[K(\mathbb{P}_{f_{\boldsymbol{\eta}}}, \mathbb{P}_{f_{\boldsymbol{\eta}'}})\big]}\right).$$

Here, $H(\cdot,\cdot)$ denotes the Hamming distance $H(\boldsymbol{\eta},\boldsymbol{\eta}') := \sum_{j=1}^{q} 1\{\eta_j \neq \eta'_j\}$, $\mathbb{P}_f$ represents the probability distribution of $(y_1,\dots,y_n)$ given $(\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)})$ when $f^* = f$, and $K(\cdot,\cdot)$ denotes the Kullback divergence between two probability distributions.

Proof of Theorem 18. Fix an integer l as

$$l = \left\lceil \frac{1}{3\log 2}\left\{ \log\left(\frac{C_{B,\bar{s}} n V^2}{\sigma^2}\right) - (\bar{s}-1)\log\log\left(\frac{C_{B,\bar{s}} n V^2}{\sigma^2}\right)\right\}\right\rceil$$

where $C_{B,\bar{s}} = B 2^{-4\bar{s}+1}(6\log 2)^{\bar{s}-1} \cdot (\bar{s}-1)!$ and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. This choice of l ensures that

$$2^{-l} \leq \left(\frac{\sigma^2}{C_{B,\bar{s}} n V^2}\right)^{1/3}\left[\log\left(\frac{C_{B,\bar{s}} n V^2}{\sigma^2}\right)\right]^{(\bar{s}-1)/3} < 2^{-l+1}. \tag{E.38}$$

Define

$$P_l = \left\{(p_1,\dots,p_{\bar{s}}) \in \mathbb{Z}^{\bar{s}}_{\geq 0} : \sum_{j=1}^{\bar{s}} p_j = l\right\},$$

and, for each $\mathbf{p} = (p_1, \dots, p_{\bar{s}}) \in P_l$, let

$$I_{\mathbf{p}} = \big\{(i_1, \dots, i_{\bar{s}}) : i_j \in [2^{p_j}] \text{ for each } j \in [\bar{s}]\big\}.$$

Recall that $[m] = \{1, \dots, m\}$ for each integer $m \geq 1$. It is clear that $|I_{\mathbf{p}}| = 2^l$ for every $\mathbf{p} \in P_l$, and

$$|P_l| = \binom{\bar{s} + l - 1}{\bar{s} - 1} \geq \frac{l^{\bar{s}-1}}{(\bar{s} - 1)!}.$$

Next, define

$$Q = \big\{(\mathbf{p}, \mathbf{i}) : \mathbf{p} \in P_l \text{ and } \mathbf{i} \in I_{\mathbf{p}}\big\}$$

and let $q = |Q| = |P_l| \cdot 2^l$. In this proof, functions will be indexed by vectors $\boldsymbol{\eta} \in \{-1, 1\}^q$, whose components are indexed by the set Q.

For an integer $m \geq 1$ and $k \in [2^m]$, denote by $\psi_{m,k}$ the real-valued function on $(0, 1)$ defined by

$$\psi_{m,k}(x) = \begin{cases} 1 & \text{if } x \in \big((k-1)2^{-m}, (k-3/4)2^{-m}\big) \cup \big((k-1/4)2^{-m}, k2^{-m}\big), \\ -1 & \text{if } x \in \big((k-3/4)2^{-m}, (k-1/4)2^{-m}\big), \\ 0 & \text{otherwise.} \end{cases}$$

Using these functions $\psi_{m,k}$, we construct $f_{\boldsymbol{\eta}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ for $\boldsymbol{\eta} \in \{-1, 1\}^q$ as follows, with which we will use Lemma 43 to prove the lower bound on the minimax risk $\mathfrak{M}^{d,s}_{n,V}$. For each $\boldsymbol{\eta} \in \{-1, 1\}^q$, let $\nu_{\boldsymbol{\eta}}$ be the signed measure on $\prod_{j=1}^{\bar{s}}(-M_j/2, M_j/2)$ defined by

$$d\nu_{\boldsymbol{\eta}}(\mathbf{t}) = \frac{1}{M_1 \cdots M_{\bar{s}}} \cdot \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \bigg(\prod_{j=1}^{\bar{s}} \psi_{p_j, i_j}\Big(\frac{t_j}{M_j} + \frac{1}{2}\Big) \bigg) \, d\mathbf{t},$$

and define $f_{\boldsymbol{\eta}} : \mathbb{R}^d \to \mathbb{R}$ as

$$f_{\boldsymbol{\eta}}(x_1, \dots, x_d) = \int_{\prod_{j=1}^{\bar{s}}(-M_j/2, M_j/2)} \prod_{j=1}^{\bar{s}} \mathbf{1}(x_j \geq t_j) \, d\nu_{\boldsymbol{\eta}}(\mathbf{t}).$$

Clearly, $f_{\boldsymbol{\eta}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ for every $\boldsymbol{\eta} \in \{-1, 1\}^q$. The following lemma, whose proof is deferred to Appendix E.20, summarizes the key properties of the functions $f_{\boldsymbol{\eta}}$ we need for the proof.

Lemma 44. *For each $\boldsymbol{\eta} \in \{-1, 1\}^q$, the complexity of $f_{\boldsymbol{\eta}}$ is bounded by V, i.e.,*

$$V^{d,s}_{\infty-xgb}(f_{\boldsymbol{\eta}}) \leq V. \tag{E.39}$$

We also have

$$\max_{H(\boldsymbol{\eta}, \boldsymbol{\eta}')=1} \|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|^2_{p_0,2} \leq \frac{BV^2}{|P_l|} \cdot 2^{-3l-4\bar{s}+2} \tag{E.40}$$

and

$$\min_{\boldsymbol{\eta} \neq \boldsymbol{\eta}'} \frac{\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|^2_{p_0,2}}{H(\boldsymbol{\eta}, \boldsymbol{\eta}')} \geq \frac{bV^2}{|P_l|} \cdot 2^{-3l-6\bar{s}+2}. \tag{E.41}$$

It is straightforward to check that the Kullback divergence between $\mathbb{P}_{f_{\boldsymbol{\eta}}}$ and $\mathbb{P}_{f_{\boldsymbol{\eta}'}}$ for $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \{-1, 1\}^q$ can be computed by

$$K(\mathbb{P}_{f_{\boldsymbol{\eta}}}, \mathbb{P}_{f_{\boldsymbol{\eta}'}}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \big(f_{\boldsymbol{\eta}}(\mathbf{x}^{(i)}) - f_{\boldsymbol{\eta}'}(\mathbf{x}^{(i)})\big)^2.$$

Hence, (E.40) gives

$$\max_{H(\boldsymbol{\eta},\boldsymbol{\eta}')=1} \mathbb{E}\big[K(\mathbb{P}_{f_{\boldsymbol{\eta}}}, \mathbb{P}_{f_{\boldsymbol{\eta}'}})\big] = \frac{n}{2\sigma^2} \cdot \max_{H(\boldsymbol{\eta},\boldsymbol{\eta}')=1} \|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2 \leq \frac{BnV^2}{\sigma^2|P_l|} \cdot 2^{-3l-4\bar{s}+1}. \tag{E.42}$$

Applying Lemma 43, along with (E.41) and (E.42), we bound the minimax risk $\mathfrak{M}_{n,V}^{d,s}$ as

$$\begin{aligned}
\mathfrak{M}_{n,V}^{d,s} &\geq \frac{q}{8} \cdot \frac{bV^2}{|P_l|} \cdot 2^{-3l-6\bar{s}+2}\left(1 - \sqrt{\frac{BnV^2}{\sigma^2|P_l|} \cdot 2^{-3l-4\bar{s}}}\right) \\
&\geq bV^2 2^{-2l-6\bar{s}-1}\left(1 - \sqrt{\frac{1}{2(6\log 2)^{\bar{s}-1}} \cdot \frac{C_{B,\bar{s}}nV^2}{\sigma^2} \cdot \frac{2^{-3l}}{l^{\bar{s}-1}}}\right).
\end{aligned}$$

Recall that $q = |P_l| \cdot 2^l$, $|P_l| \geq l^{\bar{s}-1}/(\bar{s}-1)!$, and $C_{B,\bar{s}} = B2^{-4\bar{s}+1}(6\log 2)^{\bar{s}-1} \cdot (\bar{s}-1)!$. Our choice of l (at the beginning of the proof) implies that

$$\begin{aligned}
\frac{1}{2(6\log 2)^{\bar{s}-1}} \cdot \frac{C_{B,\bar{s}}nV^2}{\sigma^2} \cdot \frac{2^{-3l}}{l^{\bar{s}-1}} &\leq \frac{1}{2} \cdot \left[\frac{\log(C_{B,\bar{s}}nV^2/\sigma^2)}{2\{\log(C_{B,\bar{s}}nV^2/\sigma^2) - (\bar{s}-1)\log\log(C_{B,\bar{s}}nV^2/\sigma^2)\}}\right]^{\bar{s}-1} \\
&\leq \frac{1}{2} \cdot \left[2\left(1 - (\bar{s}-1)\frac{\log\log(C_{B,\bar{s}}nV^2/\sigma^2)}{\log(C_{B,\bar{s}}nV^2/\sigma^2)}\right)\right]^{-(\bar{s}-1)} \\
&\leq \frac{1}{2} \cdot \left[2\left(1 - (\bar{s}-1)\Big\{\log\Big(\frac{C_{B,\bar{s}}nV^2}{\sigma^2}\Big)\Big\}^{-1/2}\right)\right]^{-(\bar{s}-1)}.
\end{aligned}$$

Here, the first inequality follows from (E.38) and

$$l \geq \frac{1}{3\log 2}\Big\{\log\Big(\frac{C_{B,\bar{s}}nV^2}{\sigma^2}\Big) - (\bar{s}-1)\log\log\Big(\frac{C_{B,\bar{s}}nV^2}{\sigma^2}\Big)\Big\},$$

and the last inequality is from the inequality $\log\log x/\log x \leq (\log x)^{-1/2}$, which is valid for all $x > 1$. If we assume that

$$n \geq \frac{e^{4\bar{s}^2}}{C_{B,\bar{s}}} \cdot \frac{\sigma^2}{V^2},$$

then

$$\frac{1}{2(6\log 2)^{\bar{s}-1}} \cdot \frac{C_{B,\bar{s}}nV^2}{\sigma^2} \cdot \frac{2^{-3l}}{l^{\bar{s}-1}} \leq \frac{1}{2} \cdot \Big(1 + \frac{1}{\bar{s}}\Big)^{-(\bar{s}-1)} \leq \frac{1}{2},$$

and thereby, we have

$$
\begin{aligned}
\mathfrak{M}_{n,V}^{d,s} &\geq \left(1-\sqrt{\frac{1}{2}}\right)\cdot bV^2 2^{-2l-6\bar{s}-1}\\
&\geq \left(1-\sqrt{\frac{1}{2}}\right)\cdot bV^2 2^{-6\bar{s}-3}\cdot\Big(\frac{\sigma^2}{C_{B,\bar{s}}nV^2}\Big)^{2/3}\Big[\log\Big(\frac{C_{B,\bar{s}}nV^2}{\sigma^2}\Big)\Big]^{2(\bar{s}-1)/3}\\
&\geq C_{b,B,\bar{s}}\Big(\frac{\sigma^2 V}{n}\Big)^{2/3}\Big[\log\Big(\frac{C_{B,\bar{s}}nV^2}{\sigma^2}\Big)\Big]^{2(\bar{s}-1)/3},
\end{aligned}
$$

where $C_{b,B,\bar{s}} = (1-1/\sqrt{2})\cdot b2^{-6\bar{s}-3}C_{B,\bar{s}}^{-2/3}$. Here, (E.38) is used again for the second inequality. Lastly, since

$$\log\Big(\frac{C_{B,\bar{s}}nV^2}{\sigma^2}\Big) \geq \frac{1}{2}\log\Big(\frac{nV^2}{\sigma^2}\Big)$$

provided that $n \geq (1/C_{B,\bar{s}}^2)\cdot(\sigma^2/V^2)$, by further assuming that

$$n \geq \max\Big(\frac{e^{4\bar{s}^2}}{C_{B,\bar{s}}}\cdot\frac{\sigma^2}{V^2}, \frac{1}{C_{B,\bar{s}}^2}\cdot\frac{\sigma^2}{V^2}\Big),$$

we can derive the lower bound

$$\mathfrak{M}_{n,V}^{d,s} \geq C'_{b,B,\bar{s}}\Big(\frac{\sigma^2 V}{n}\Big)^{2/3}\Big[\log\Big(\frac{nV^2}{\sigma^2}\Big)\Big]^{2(\bar{s}-1)/3},$$

where $C'_{b,B,\bar{s}} = C_{b,B,\bar{s}}\cdot 2^{-2(\bar{s}-1)/3}$. □

E.15 Proof of Proposition 14

Proof of Proposition 14. Suppose $f_{\mathbf{a}} \in \mathcal{F}_{\infty-\mathrm{st}}^{d,s}$ for $\mathbf{a} \in \{-\infty,+\infty\}^d$ and $\sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} f_{\mathbf{a}} \equiv f$. By repeating the argument in the proof of Proposition 12, it can be shown that for every $g \in \mathcal{F}_{\infty-\mathrm{st}}^{d,s}$,

$$\widetilde{V}_{\infty-\mathrm{xgb}}^{d,s}(g) \leq V_{\mathbf{a}}(g) = \mathrm{HK}_{\mathbf{a}}(g),$$

where $V_{\mathbf{a}}(\cdot)$ is defined as in (E.20). Applying this inequality to each $f_{\mathbf{a}} \in \mathcal{F}_{\infty-\mathrm{st}}^{d,s}$ yields

$$\sum_{\mathbf{a}\in\{-\infty,+\infty\}^d}\mathrm{HK}_{\mathbf{a}}(f_{\mathbf{a}}) \geq \sum_{\mathbf{a}\in\{-\infty,+\infty\}^d}\widetilde{V}_{\infty-\mathrm{xgb}}^{d,s}(f_{\mathbf{a}}) \geq \widetilde{V}_{\infty-\mathrm{xgb}}^{d,s}(f),$$

which proves one direction of the desired identity:

$$\widetilde{V}_{\infty-\mathrm{xgb}}^{d,s}(f) \leq \inf\Big\{\sum_{\mathbf{a}\in\{-\infty,+\infty\}^d}\mathrm{HK}_{\mathbf{a}}(f_{\mathbf{a}}) : \sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} f_{\mathbf{a}} \equiv f,\ f_{\mathbf{a}} \in \mathcal{F}_{\infty-\mathrm{st}}^{d,s}\ \forall\mathbf{a}\Big\}.$$

We now turn to the reverse inequality. Suppose $f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ is expressed as

$$f(x_1, \ldots, x_d) = c + \sum_{\substack{L,U: L\cap U=\emptyset \\ 0<|L|+|U|\leq s}} \int_{\mathbb{R}^{|L|+|U|}} \prod_{j\in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j\in U} \mathbf{1}(x_j < u_j)\, d\nu_{L,U}(\mathbf{l}, \mathbf{u})$$

for some $c \in \mathbb{R}$ and signed measures $\nu_{L,U}$ on $\mathbb{R}^{|L|+|U|}$. For each $\mathbf{a} \in \{-\infty, +\infty\}^d$, define $f_{\mathbf{a}} : \mathbb{R}^d \to \mathbb{R}$ by

$$\begin{aligned} f_{\mathbf{a}}(x_1, \ldots, x_d) = c \cdot \prod_{j=1}^{d} \mathbf{1}(a_j = -\infty) + \sum_{\substack{L,U: L\cap U=\emptyset \\ 0<|L|+|U|\leq s}} \prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty) \\ \cdot \int_{\mathbb{R}^{|L|+|U|}} \prod_{j\in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j\in U} \mathbf{1}(x_j < u_j)\, d\nu_{L,U}(\mathbf{l}, \mathbf{u}). \end{aligned}$$

It is clear that $f_{\mathbf{a}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ for all $\mathbf{a} \in \{-\infty, +\infty\}^d$. Moreover, since for each integral over $\nu_{L,U}$, there is exactly one value of $\mathbf{a}$ that makes all multiplied indicator functions equal to one, we have

$$\sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} f_{\mathbf{a}} \equiv f.$$

Fix $\mathbf{a} \in \{-\infty, +\infty\}^d$. For each nonempty $S \subseteq [d]$, we have

$$\begin{aligned} f^{S}_{(a_j, j\in S^c)}(x_j, j \in S) &= c \cdot \prod_{j=1}^{d} \mathbf{1}(a_j = -\infty) \\ &\quad + \sum_{\substack{L,U\subseteq S: L\cap U=\emptyset \\ 0<|L|+|U|\leq s}} \prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty) \\ &\qquad \cdot \int_{\mathbb{R}^{|L|+|U|}} \prod_{j\in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j\in U} \mathbf{1}(x_j < u_j)\, d\nu_{L,U}(\mathbf{l}, \mathbf{u}). \end{aligned}$$

Hence, for each nonempty $T \subseteq [d]$,

$$\begin{aligned} &\sum_{S\subseteq T} (-1)^{|T|-|S|} \cdot f^{S}_{(a_j, j\in S^c)}(x_j, j \in S) \\ &= \sum_{S\subseteq T} (-1)^{|T|-|S|} \sum_{\substack{L,U\subseteq S: L\cap U=\emptyset \\ 0<|L|+|U|\leq s}} \prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty) \\ &\qquad \cdot \int_{\mathbb{R}^{|L|+|U|}} \prod_{j\in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j\in U} \mathbf{1}(x_j < u_j)\, d\nu_{L,U}(\mathbf{l}, \mathbf{u}) \\ &= \sum_{\substack{L,U: L\cap U=\emptyset \\ 0<|L|+|U|\leq s}} \Bigg(\sum_{S: L\cup U\subseteq S\subseteq T} (-1)^{|T|-|S|} \Bigg) \prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty) \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad \cdot \int_{\mathbb{R}^{|L|+|U|}} \prod_{j\in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j\in U} \mathbf{1}(x_j < u_j)\, d\nu_{L,U}(\mathbf{l},\mathbf{u}) \\
&= \sum_{\substack{L,U: L\cap U=\emptyset, L\cup U = T\\ 0<|L|+|U|\leq s}} \prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty) \\
&\qquad\qquad\qquad \cdot \int_{\mathbb{R}^{|L|+|U|}} \prod_{j\in L} \mathbf{1}(x_j \geq l_j) \cdot \prod_{j\in U} \mathbf{1}(x_j < u_j)\, d\nu_{L,U}(\mathbf{l},\mathbf{u})
\end{aligned}
$$

if $|T| \leq s$, and it vanishes otherwise. For each nonempty $T \subseteq [d]$ with $|T| \leq s$, since there is at most one pair of (L, U) with $L \cap U = \emptyset$ and $L \cup U = T$ such that

$$
\prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty) = 1,
$$

by repeating the computation in the proof of Proposition 12, we obtain

$$
\begin{aligned}
\mathrm{Vit}(f^T_{(a_j, j\in T^c)}) &= \mathrm{Vit}\Big((x_j, j \in T) \mapsto \sum_{S\subseteq T} (-1)^{|T|-|S|} \cdot f^S_{(a_j, j\in S^c)}(x_j, j \in S)\Big) \\
&= \sum_{\substack{L,U: L\cap U=\emptyset, L\cup U = T\\ 0<|L|+|U|\leq s}} \prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty) \cdot \|\nu_{L,U}\|_{\mathrm{TV}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathrm{HK}_{\mathbf{a}}(f_{\mathbf{a}}) &= \sum_{0<|T|\leq d} \mathrm{Vit}(f^T_{(a_j, j\in T^c)}) = \sum_{0<|T|\leq s} \mathrm{Vit}(f^T_{(a_j, j\in T^c)}) \\
&= \sum_{0<|T|\leq s} \sum_{\substack{L,U: L\cap U=\emptyset, L\cup U = T\\ 0<|L|+|U|\leq s}} \prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty) \cdot \|\nu_{L,U}\|_{\mathrm{TV}} \\
&= \sum_{\substack{L,U: L\cap U=\emptyset\\ 0<|L|+|U|\leq s}} \prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty) \cdot \|\nu_{L,U}\|_{\mathrm{TV}}.
\end{aligned}
$$

Summing the above identity over all $\mathbf{a} \in \{-\infty, +\infty\}^d$ gives

$$
\begin{aligned}
&\inf\Big\{ \sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} \mathrm{HK}_{\mathbf{a}}(f_{\mathbf{a}}) : \sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} f_{\mathbf{a}} \equiv f,\ f_{\mathbf{a}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}\ \forall \mathbf{a}\Big\} \leq \sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} \mathrm{HK}_{\mathbf{a}}(f_{\mathbf{a}}) \\
&\quad = \sum_{\substack{L,U: L\cap U=\emptyset\\ 0<|L|+|U|\leq s}} \sum_{\mathbf{a}\in\{-\infty,+\infty\}^d} \Big(\prod_{j\in U^c} \mathbf{1}(a_j = -\infty) \cdot \prod_{j\in U} \mathbf{1}(a_j = +\infty)\Big) \cdot \|\nu_{L,U}\|_{\mathrm{TV}} \\
&\quad = \sum_{\substack{L,U: L\cap U=\emptyset\\ 0<|L|+|U|\leq s}} \|\nu_{L,U}\|_{\mathrm{TV}}.
\end{aligned}
$$

Taking the infimum over all possible representations $f_{c,\{\nu_{L,U}\}}$ of f, we arrive at

$$\inf\Big\{\sum_{\mathbf{a}\in\{-\infty,+\infty\}^d}\mathrm{HK}_{\mathbf{a}}(f_{\mathbf{a}}):\sum_{\mathbf{a}\in\{-\infty,+\infty\}^d}f_{\mathbf{a}}\equiv f,\ f_{\mathbf{a}}\in\mathcal{F}^{d,s}_{\infty-\mathrm{st}}\ \forall\mathbf{a}\Big\}\le\widetilde{V}^{d,s}_{\infty-\mathrm{xgb}}(f),$$

which completes the proof. □

E.16 Proof of Lemma 7

Proof of Lemma 7. Suppose that $f\in\mathcal{F}^{d,s}_{\mathrm{st}}$ and that $f=\sum_{k=1}^K f_k$, where each f_k is a regression tree with right-continuous splits and depth at most s. Let $\mathbf{w}_k$ denote the leaf-weight vector of f_k.

Fix an integer $L\ge 1$. For each k, define the regression tree $g_{k,L}:=(1/L)f_k$, obtained by scaling each leaf weight of f_k by $1/L$ while keeping the same tree structure. Using these $g_{k,L}$, we can represent f as a sum of KL trees:

$$f=\sum_{k=1}^K\sum_{l=1}^L g_{k,L}.$$

For this representation, the sum of the p^{th} powers of the leaf weights equals

$$\sum_{k=1}^K\sum_{l=1}^L\|(1/L)\cdot\mathbf{w}_k\|_p^p=\frac{1}{L^{p-1}}\sum_{k=1}^K\|\mathbf{w}_k\|_p^p.$$

Because $p>1$, this quantity converges to 0 as $L\to\infty$. This proves that $V^{d,s}_{\mathrm{xgb}}(f;p)=0$. □

E.17 Proof of Lemma 40

Lemma 40 is proved similarly to Lemma 16 in Appendix C. It is also a corollary of the more general result involving the Bernstein norm stated in Lemma 23 of Appendix C.

Proof of Lemma 40. Let P be the law of $(\mathbf{x}^{(i)},\epsilon_i)$ on $\mathbb{R}^d\times\{-1,1\}$ and let $\mathcal{G}$ denote the collection of all functions Φ_f on $\mathbb{R}^d\times\{-1,1\}$, one for each $f\in\mathcal{F}$, defined by

$$\Phi_f(\mathbf{x},\epsilon)=\frac{\epsilon f(\mathbf{x})}{2D}\quad\text{for }(\mathbf{x},\epsilon)\in\mathbb{R}^d\times\{-1,1\}.$$

For every $\Phi_f\in\mathcal{G}$, we have

$$\|\Phi_f\|_{P,B}=\left(2\mathbb{E}_P\left[\exp\left(\left|\frac{\epsilon f(\mathbf{x})}{2D}\right|\right)-1-\left|\frac{\epsilon f(\mathbf{x})}{2D}\right|\right]\right)^{1/2}=\left(2\sum_{m=2}^\infty\frac{1}{m!}\cdot\mathbb{E}_{\mathbf{x}\sim p_0}\left[\left|\frac{f(\mathbf{x})}{2D}\right|^m\right]\right)^{1/2}$$

$$\leq \left(2\sum_{m=2}^{\infty}\frac{1}{m!}\cdot\mathbb{E}_{\mathbf{x}\sim p_0}\Big[\Big|\frac{f(\mathbf{x})}{2D}\Big|^2\Big]\right)^{1/2} \leq \frac{(e-2)^{1/2}t}{2^{1/2}D} =: \frac{at}{2D},$$

where the first inequality uses the fact that $\|f\|_\infty \leq D$, and the second inequality follows from the fact that $\|f\|_{p_0,2} \leq t$. Applying Lemma 23 with $\mathcal{G}$ and $\delta = at/2D$, we obtain

$$\begin{aligned}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\epsilon_i f(\mathbf{x}^{(i)})\Big|\Big] &= 2D\cdot\mathbb{E}\Big[\sup_{\Phi_f\in\mathcal{G}}\Big|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\Phi_f(\mathbf{x}^{(i)},\epsilon_i)\Big|\Big]\\ &\leq 2D\cdot CJ_{[\,]}\Big(\frac{at}{2D},\mathcal{G},\|\cdot\|_{P,B}\Big)\cdot\Big(1+\frac{J_{[\,]}(at/(2D),\mathcal{G},\|\cdot\|_{P,B})}{(at/(2D))^2\sqrt{k}}\Big).\end{aligned} \tag{E.43}$$

Next, we relate the bracketing entropy integral of $\mathcal{G}$ in the Bernstein norm to that of $\mathcal{F}$ in the $\|\cdot\|_{p_0,2}$ norm. Fix $\Phi_f \in \mathcal{G}$, and let $[f_1, f_2]$ be a bracket containing f. Since $\|f\|_\infty \leq D$, by replacing f_1 with $\mathbf{x} \mapsto \max(\min(f_1(\mathbf{x}), D), -D)$ if necessary, we assume that $\|f_1\|_\infty \leq D$. Similarly, we assume that $\|f_2\|_\infty \leq D$. Define $\Phi_1, \Phi_2 : \mathbb{R}^d \times \{-1, 1\} \to \mathbb{R}$ by

$$\Phi_1(\mathbf{x},\epsilon) = f_1(\mathbf{x})\cdot\mathbf{1}(\epsilon=1) - f_2(\mathbf{x})\cdot\mathbf{1}(\epsilon=-1) \quad \text{for } (\mathbf{x},\epsilon)\in\mathbb{R}^d\times\{-1,1\}$$

and

$$\Phi_2(\mathbf{x},\epsilon) = f_2(\mathbf{x})\cdot\mathbf{1}(\epsilon=1) - f_1(\mathbf{x})\cdot\mathbf{1}(\epsilon=-1) \quad \text{for } (\mathbf{x},\epsilon)\in\mathbb{R}^d\times\{-1,1\}.$$

Clearly, the bracket $[\Phi_1, \Phi_2]$ contains Φ_f. Moreover, since $\|f_2 - f_1\|_\infty \leq 2D$, by the same argument as above, we obtain

$$\|\Phi_2-\Phi_1\|_{P,B} = \left(2\mathbb{E}_{\mathbf{x}\sim p_0}\Big[\exp\Big(\Big|\frac{(f_2-f_1)(\mathbf{x})}{2D}\Big|\Big)-1-\Big|\frac{(f_2-f_1)(\mathbf{x})}{2D}\Big|\Big]\right)^{1/2} \leq \frac{a}{2D}\cdot\|f_2-f_1\|_{p_0,2}.$$

It follows that

$$N_{[\,]}\Big(\frac{a\epsilon}{2D},\mathcal{G},\|\cdot\|_{P,B}\Big) \leq N_{[\,]}(\epsilon,\mathcal{F},\|\cdot\|_{p_0,2}) \quad \text{for } \epsilon > 0.$$

Hence,

$$\begin{aligned}J_{[\,]}\Big(\frac{at}{2D},\mathcal{G},\|\cdot\|_{P,B}\Big) &= \int_0^{at/2D}\sqrt{1+N_{[\,]}(\epsilon,\mathcal{G},\|\cdot\|_{P,B})}\,d\epsilon\\ &= \frac{a}{2D}\int_0^t\sqrt{1+N_{[\,]}\Big(\frac{a\epsilon}{2D},\mathcal{G},\|\cdot\|_{P,B}\Big)}\,d\epsilon \leq \frac{a}{2D}\cdot J_{[\,]}(t,\mathcal{F},\|\cdot\|_{p_0,2}).\end{aligned}$$

Substituting this bound into (E.43) completes the proof. □

E.18 Proof of Lemma 41

We use the following lemma, which ensures that the supports of the signed measures can be restricted to $\prod_{j\in L}(-M_j/2, M_j/2] \times \prod_{j\in U}(-M_j/2, M_j/2]$ without changing the function on $\prod_{j=1}^{d}[-M_j/2, M_j/2]$. The proof is omitted, as it can be proved similarly to Lemma 39.

Lemma 45. *For every $f_{c,\{\nu_{L,U}\}}$, there exists $f_{b,\{\mu_{L,U}\}}$ with finite signed measures $\mu_{L,U}$ supported on $\prod_{j\in L}(-M_j/2, M_j/2] \times \prod_{j\in U}(-M_j/2, M_j/2]$ such that*

- $f_{b,\{\mu_{L,U}\}}(\cdot) = f_{c,\{\nu_{L,U}\}}(\cdot)$ *on* $\prod_{j=1}^{d}[-M_j/2, M_j/2]$;
- $\sum_{0<|L|+|U|\leq s} \|\mu_{L,U}\|_{TV} \leq \sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{TV}$.

Proof of Lemma 41. Suppose $f \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ with $V^{d,s}_{\infty-\mathrm{xgb}}(f) < V$. By the definition of the complexity measure $V^{d,s}_{\infty-\mathrm{xgb}}(\cdot)$, there exists $f_{c,\{\nu_{L,U}\}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{st}}$ such that $f_{c,\{\nu_{L,U}\}} \equiv f$ and

$$\sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{\mathrm{TV}} \leq V.$$

Lemma 45 also guarantees the existence of $f_{b,\{\mu_{L,U}\}}$ with finite signed measures $\mu_{L,U}$ supported on $\prod_{j\in L}(-M_j/2, M_j/2] \times \prod_{j\in U}(-M_j/2, M_j/2]$ satisfying conditions (a) and (b) of the lemma. By condition (a), $f_{b,\{\mu_{L,U}\}}(\cdot) = f_{c,\{\nu_{L,U}\}}(\cdot) = f(\cdot)$ on $\prod_{j=1}^{d}[-M_j/2, M_j/2]$. Moreover, by condition (b),

$$\sum_{0<|L|+|U|\leq s} \|\mu_{L,U}\|_{\mathrm{TV}} \leq \sum_{0<|L|+|U|\leq s} \|\nu_{L,U}\|_{\mathrm{TV}} \leq V.$$

Hence, $f_{b,\{\mu_{L,U}\}}$ is a desired function satisfying the conditions of Lemma 41. □

E.19 Proof of Lemma 42

Proof of Lemma 42. We use the bracketing entropy bound for $D_{\mathbf{M}}(V,t)$ established in Lemma 6:

$$\begin{aligned} J_{[\,]}(t, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2}) &= \int_0^t \sqrt{1 + \log N_{[\,]}(\epsilon, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2})}\, d\epsilon \\ &\leq t + \int_0^t \left[\log\left(2 + \frac{C_s(V+t)}{\epsilon}\right)\right]^{1/2} d\epsilon \\ &\quad + C_{B,s} d^{\bar{s}} (1+\log d)^{\bar{s}-1} \int_0^t \left(2 + \frac{V}{\epsilon}\right)^{1/2} \left[\log\left(2 + \frac{V}{\epsilon}\right)\right]^{\bar{s}-1} d\epsilon. \end{aligned}$$

To handle the integrals on the right-hand side, we use the following lemma, which is a straightforward consequence of integration by parts. We omit the proof, since it is similar to that of Lemma 13 in Appendix C.

Lemma 46. *For $u > t$,*

$$\int_0^t \left[\log\left(\frac{u}{\epsilon}\right)\right]^{1/2} d\epsilon = \frac{t}{2\sqrt{\tau}} \cdot (1 + 2\tau)$$

and

$$\int_0^t \left(\frac{u}{\epsilon}\right)^{1/2} \left[\log\left(\frac{u}{\epsilon}\right)\right]^{k} d\epsilon \leq C_k u^{1/2} t^{1/2} (1 + \tau^k),$$

where $\tau = \log(u/t)$ and C_k is a constant depending on k.

By the first inequality in Lemma 46,

$$\int_0^t \Big[\log\Big(2+\frac{C_s(V+t)}{\epsilon}\Big)\Big]^{1/2} d\epsilon \le \int_0^t \Big[\log\Big(\frac{2t+C_s(V+t)}{\epsilon}\Big)\Big]^{1/2} d\epsilon$$
$$\le Ct\Big[1+2\log\Big(\frac{2t+C_s(V+t)}{t}\Big)\Big] \le C_s t\log\Big(2+\frac{V}{t}\Big).$$

Also, by the second inequality in Lemma 46 and by $(x+y)^{1/2} \le x^{1/2}+y^{1/2}$, we have

$$\int_0^t \Big(2+\frac{V}{\epsilon}\Big)^{1/2}\Big[\log\Big(2+\frac{V}{\epsilon}\Big)\Big]^{\bar{s}-1} d\epsilon \le \int_0^t \Big(\frac{2t+V}{\epsilon}\Big)^{1/2}\Big[\log\Big(\frac{2t+V}{\epsilon}\Big)\Big]^{\bar{s}-1} d\epsilon$$
$$\le C_s(2t+V)^{1/2}t^{1/2}\Big(1+\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1}\Big)$$
$$\le C_s t\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1} + C_s V^{1/2}t^{1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1}.$$

Combining these bounds yields

$$J_{[\,]}(t, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2}) \le C_s t\log\Big(2+\frac{V}{t}\Big) + C_{B,s}d^{\bar{s}}(1+\log d)^{\bar{s}-1}t\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1}$$
$$+ C_{B,s}d^{\bar{s}}(1+\log d)^{\bar{s}-1}V^{1/2}t^{1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1}.$$

If $t \le V$,

$$t\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1} \le V^{1/2}t^{1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1},$$

and otherwise,

$$t\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1} \le C_s t \le C_s t\log\Big(2+\frac{V}{t}\Big).$$

Hence, in both cases, we have

$$J_{[\,]}(t, D_{\mathbf{M}}(V,t), \|\cdot\|_{p_0,2}) \le C_{B,s}d^{\bar{s}}(1+\log d)^{\bar{s}-1}\Big(t\log\Big(2+\frac{V}{t}\Big) + V^{1/2}t^{1/2}\Big[\log\Big(2+\frac{V}{t}\Big)\Big]^{\bar{s}-1}\Big).$$

□

E.20 Proof of Lemma 44

Proof of (E.39). By definition, for each $\boldsymbol{\eta} \in \{-1,1\}^q$, we have

$$|\nu_{\boldsymbol{\eta}}|\Big(\prod_{j=1}^{\bar{s}}\Big(-\frac{M_j}{2}, \frac{M_j}{2}\Big)\Big)$$

$$
\begin{aligned}
&= \frac{1}{M_1 \cdots M_{\bar{s}}} \cdot \frac{V}{\sqrt{|P_l|}} \int_{\prod_{j=1}^{\bar{s}}(-M_j/2, M_j/2)} \left| \sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^{\bar{s}} \psi_{p_j,i_j}\Big(\frac{t_j}{M_j} + \frac{1}{2}\Big) \right| d\mathbf{t} \\
&= \frac{V}{\sqrt{|P_l|}} \int_{(0,1)^{\bar{s}}} \left| \sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^{\bar{s}} \psi_{p_j,i_j}(t_j) \right| d\mathbf{t} \\
&\leq \frac{V}{\sqrt{|P_l|}} \left(\int_{(0,1)^{\bar{s}}} \Big(\sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^{\bar{s}} \psi_{p_j,i_j}(t_j) \Big)^2 d\mathbf{t} \right)^{1/2} \\
&= \frac{V}{\sqrt{|P_l|}} \left(\int_{(0,1)^{\bar{s}}} \sum_{\mathbf{p} \in P_l} \Big(\sum_{\mathbf{i} \in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^{\bar{s}} \psi_{p_j,i_j}(t_j) \Big)^2 d\mathbf{t} \right)^{1/2} \\
&= \frac{V}{\sqrt{|P_l|}} \left(\sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \int_{(0,1)^{\bar{s}}} \Big(\prod_{j=1}^{\bar{s}} \psi_{p_j,i_j}(t_j) \Big)^2 d\mathbf{t} \right)^{1/2} \\
&= \frac{V}{\sqrt{|P_l|}} \left(\sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \prod_{j=1}^{\bar{s}} \int_0^1 (\psi_{p_j,i_j}(t_j))^2 \, dt_j \right)^{1/2} = \frac{V}{\sqrt{|P_l|}} \left(\sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \prod_{j=1}^{\bar{s}} 2^{-p_j} \right)^{1/2} = V.
\end{aligned}
$$

Here, the inequality is from Cauchy inequality, the third equality follows from the fact that

$$
\int_0^1 \psi_{m,k}(x) \psi_{m',k'}(x) \, dx = 0
$$

for distinct m and m', and the fourth equality is due to the fact that $\psi_{m,k}\psi_{m,k'} \equiv 0$ provided $k \neq k'$. This proves that $V^{d,s}_{\infty-\mathrm{xgb}}(f_{\boldsymbol{\eta}}) \leq V$ for every $\boldsymbol{\eta} \in \{-1,1\}^q$. $\square$

Proof of (E.40). For an integer $m \geq 1$ and $k \in [2^m]$, let $\Psi_{m,k}$ be the real-valued function on $[0,1]$ defined by

$$
\Psi_{m,k}(x) = \int_0^x \psi_{m,k}(t) \, dt.
$$

It can be readily verified that

$$
\begin{aligned}
&\text{(i) } \Psi_{m,k}(x) = 0 \ \text{ if } x \leq (k-1)2^{-m} \text{ or } x \geq k2^{-m} \\
&\text{(ii) } \Psi_{m,k}(x + 2^{-m-1}) = -\Psi_{m,k}(x) \ \text{ for } x \in [(k-1)2^{-m}, (k-1/2)2^{-m}] \\
&\text{(iii) } |\Psi_{m,k}(x)| \leq 2^{-m-2} \ \text{ for all } x \in [0,1].
\end{aligned}
\tag{E.44}
$$

Also, for every $(x_1, \dots, x_{\bar{s}}) \in [0,1]^{\bar{s}}$, we have

$$
\begin{aligned}
&f_{\boldsymbol{\eta}}\Big(M_1 x_1 - \frac{M_1}{2}, \dots, M_{\bar{s}} x_{\bar{s}} - \frac{M_{\bar{s}}}{2}\Big) \\
&\quad = \frac{1}{M_1 \cdots M_{\bar{s}}} \cdot \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \eta_{\mathbf{p},\mathbf{i}} \int_{\prod_{j=1}^{\bar{s}}(-M_j/2, M_j/2)} \prod_{j=1}^{\bar{s}} \Big[\mathbf{1}\Big(M_j x_j - \frac{M_j}{2} \geq t_j \Big)
\end{aligned}
\tag{E.45}
$$

$$\cdot \psi_{p_j,i_j}\Big(\frac{t_j}{M_j}+\frac{1}{2}\Big)\Big]\,d\mathbf{t}$$

$$= \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l}\sum_{\mathbf{i}\in I_\mathbf{p}} \eta_{\mathbf{p},\mathbf{i}} \cdot \int_{(0,1)^{\bar{s}}} \prod_{j=1}^{\bar{s}} \big[\mathbf{1}(x_j \geq t_j)\cdot \psi_{p_j,i_j}(t_j)\big]\,d\mathbf{t}$$

$$= \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l}\sum_{\mathbf{i}\in I_\mathbf{p}} \eta_{\mathbf{p},\mathbf{i}} \prod_{j=1}^{\bar{s}} \int_0^{x_j} \psi_{p_j,i_j}(t_j)\,dt_j = \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l}\sum_{\mathbf{i}\in I_\mathbf{p}} \eta_{\mathbf{p},\mathbf{i}} \cdot \prod_{j=1}^{\bar{s}} \Psi_{p_j,i_j}(x_j). \quad \text{(E.46)}$$

We prove (E.40) using equation (E.45). Assume that we are given $\boldsymbol{\eta},\boldsymbol{\eta}' \in \{-1,1\}^q$ with $H(\boldsymbol{\eta},\boldsymbol{\eta}')=1$ and that $(\mathbf{p},\mathbf{i})$ is the unique element in Q for which $\eta_{\mathbf{p},\mathbf{i}} \neq \eta'_{\mathbf{p},\mathbf{i}}$. We then have

$$(f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'})\Big(M_1x_1 - \frac{M_1}{2},\ldots,M_{\bar{s}}x_{\bar{s}} - \frac{M_{\bar{s}}}{2}\Big) = \frac{V}{\sqrt{|P_l|}}\cdot(\eta_{\mathbf{p},\mathbf{i}} - \eta'_{\mathbf{p},\mathbf{i}})\prod_{j=1}^{\bar{s}} \Psi_{p_j,i_j}(x_j),$$

from which it follows that

$$\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|^2_{p_0,2} \leq \frac{B}{M_1\cdots M_d}\cdot \int_{\prod_{j=1}^d[-M_j/2,M_j/2]} (f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'})^2(x_1,\ldots,x_{\bar{s}})\,d\mathbf{x}$$

$$= B\cdot \int_{[0,1]^{\bar{s}}} (f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'})^2\Big(M_1x_1 - \frac{M_1}{2},\ldots,M_{\bar{s}}x_{\bar{s}} - \frac{M_{\bar{s}}}{2}\Big)\,d\mathbf{x}$$

$$= \frac{4BV^2}{|P_l|}\cdot \prod_{j=1}^{\bar{s}} \int_0^1 \Psi^2_{p_j,i_j}(x_j)\,dx_j \leq \frac{4BV^2}{|P_l|}\cdot \prod_{j=1}^{\bar{s}} 2^{-p_j}\cdot 2^{-2p_j-4} = \frac{BV^2}{|P_l|}\cdot 2^{-3l-4\bar{s}+2}.$$

Recall that $B = M_1\cdots M_d \cdot \sup_\mathbf{x} p_0(\mathbf{x})$ for the first inequality. □

Proof of (E.41). Fix $\boldsymbol{\eta}\neq\boldsymbol{\eta}' \in \{-1,1\}^q$. For an integer $m\geq 1$ and $k\in[2^m]$, let $h_{m,k}$ be the real-valued function on $[0,1]$ defined by

$$h_{m,k}(x) = \begin{cases} 2^{m/2} & \text{if } (k-1)2^{-m} < x < (k-1/2)2^{-m}, \\ -2^{m/2} & \text{if } (k-1/2)2^{-m} < x < k2^{-m}, \\ 0 & \text{otherwise,} \end{cases}$$

and, for each $(\mathbf{p},\mathbf{i})\in Q$, let $H_{\mathbf{p},\mathbf{i}}$ be the real-valued function on $[0,1]^{\bar{s}}$ defined by

$$H_{\mathbf{p},\mathbf{i}}(x_1,\ldots,x_{\bar{s}}) = \prod_{j=1}^{\bar{s}} h_{p_j,i_j}(x_j).$$

It can be readily checked that $\{H_{\mathbf{p},\mathbf{i}} : (\mathbf{p},\mathbf{i})\in Q\}$ is an orthonormal set in $L^2([0,1]^{\bar{s}})$. Consider the function $g_{\boldsymbol{\eta},\boldsymbol{\eta}'} : [0,1]^{\bar{s}} \to \mathbb{R}$ defined by

$$g_{\boldsymbol{\eta},\boldsymbol{\eta}'}(x_1,\ldots,x_{\bar{s}}) = \frac{V}{\sqrt{|P_l|}} \sum_{\mathbf{p}\in P_l}\sum_{\mathbf{i}\in I_\mathbf{p}} \big(\eta_{\mathbf{p},\mathbf{i}} - \eta'_{\mathbf{p},\mathbf{i}}\big) \prod_{j=1}^{\bar{s}} \Psi_{p_j,i_j}(x_j).$$

Since

$$b = M_1 \cdots M_d \cdot \inf_{\mathbf{x} \in \prod_{j=1}^d [-M_j/2, M_j/2]} p_0(\mathbf{x}) > 0,$$

we have

$$\begin{aligned} \|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2 &\geq \frac{b}{M_1 \cdots M_d} \cdot \int_{\prod_{j=1}^d [-M_j/2, M_j/2]} (f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'})^2 (x_1, \ldots, x_{\bar{s}}) \, d\mathbf{x} \\ &= b \cdot \int_{[0,1]^{\bar{s}}} (f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'})^2 \Big(M_1 x_1 - \frac{M_1}{2}, \ldots, M_{\bar{s}} x_{\bar{s}} - \frac{M_{\bar{s}}}{2} \Big) \, d\mathbf{x} = b \|g_{\boldsymbol{\eta},\boldsymbol{\eta}'}\|_2^2, \end{aligned}$$

where $\|\cdot\|_2$ denotes the L^2 norm. Recall (E.45) in the proof of (E.40) for the last equality. By Bessel's inequality, it thus follows that

$$\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2 \geq b \|g_{\boldsymbol{\eta},\boldsymbol{\eta}'}\|_2^2 \geq b \sum_{\mathbf{p}' \in P_l} \sum_{\mathbf{i}' \in I_{\mathbf{p}'}} \langle g_{\boldsymbol{\eta},\boldsymbol{\eta}'}, H_{\mathbf{p}',\mathbf{i}'} \rangle^2, \tag{E.47}$$

where $\langle \cdot, \cdot \rangle$ denotes the L^2 inner product.

Observe that for each $(\mathbf{p}', \mathbf{i}') \in Q$,

$$\langle g_{\boldsymbol{\eta},\boldsymbol{\eta}'}, H_{\mathbf{p}',\mathbf{i}'} \rangle = \frac{V}{\sqrt{|P_l|}} \cdot \sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} \left(\eta_{\mathbf{p},\mathbf{i}} - \eta'_{\mathbf{p},\mathbf{i}} \right) \prod_{j=1}^{\bar{s}} \langle \Psi_{p_j,i_j}, h_{p'_j,i'_j} \rangle. \tag{E.48}$$

We conclude the proof by showing that for $(\mathbf{p}, \mathbf{i}), (\mathbf{p}', \mathbf{i}') \in Q$,

$$\prod_{j=1}^{\bar{s}} \langle \Psi_{p_j,i_j}, h_{p'_j,i'_j} \rangle = \begin{cases} 2^{-3l/2 - 3\bar{s}} & \text{if } (\mathbf{p}, \mathbf{i}) = (\mathbf{p}', \mathbf{i}') \\ 0 & \text{otherwise.} \end{cases} \tag{E.49}$$

Once (E.49) is proved, by combining it with (E.47) and (E.48), we can derive that

$$\|f_{\boldsymbol{\eta}} - f_{\boldsymbol{\eta}'}\|_{p_0,2}^2 \geq \frac{bV^2}{|P_l|} \cdot 2^{-3l - 6\bar{s}} \sum_{\mathbf{p} \in P_l} \sum_{\mathbf{i} \in I_{\mathbf{p}}} (\eta_{\mathbf{p},\mathbf{i}} - \eta'_{\mathbf{p},\mathbf{i}})^2 = \frac{bV^2}{|P_l|} \cdot 2^{-3l - 6\bar{s} + 2} \cdot H(\boldsymbol{\eta}, \boldsymbol{\eta}'),$$

from which (E.41) directly follows. We first consider the case where $(\mathbf{p}, \mathbf{i}) \neq (\mathbf{p}', \mathbf{i}')$. If $\mathbf{p} \neq \mathbf{p}'$, then, since $\sum_{j=1}^{\bar{s}} p_j = l = \sum_{j=1}^{\bar{s}} p'_j$, there exists $j \in [\bar{s}]$ such that $p_j > p'_j$. In this case, $h_{p'_j,i'_j}$ is constant on $((i_j - 1)2^{-p_j}, i_j 2^{-p_j})$, and hence, (E.44) implies that $\langle \Psi_{p_j,i_j}, h_{p'_j,i'_j} \rangle = 0$. If $\mathbf{p} = \mathbf{p}'$, then $\mathbf{i}$ and $\mathbf{i}'$ must be distinct, and thus, there exists $j \in [\bar{s}]$ such that $i_j \neq i'_j$. In this case, $\Psi_{p_j,i_j}(x) \cdot h_{p'_j,i'_j}(x) = 0$ for all $x \in [0, 1]$, and clearly, $\langle \Psi_{p_j,i_j}, h_{p'_j,i'_j} \rangle = 0$. For the case where $(\mathbf{p}, \mathbf{i}) = (\mathbf{p}', \mathbf{i}')$, (E.49) follows from the fact that for each $j \in [\bar{s}]$,

$$\langle \Psi_{p_j,i_j}, h_{p_j,i_j} \rangle = \int_{(i_j-1)2^{-p_j}}^{i_j 2^{-p_j}} \Psi_{p_j,i_j}(x) \cdot h_{p_j,i_j}(x) \, dx = 2^{-3p_j/2 - 3}.$$

$\square$