The Annals of Statistics
2024, Vol. 52, No. 3, 1102–1126
https://doi.org/10.1214/24-AOS2384

MARS VIA LASSO

BY DOHYEONG KI[1,a], BILLY FANG[2,c] AND ADITYANAND GUNTUBOYINA[1,b]

[1]*Department of Statistics, University of California, Berkeley,* [a]*dohyeong_ki@berkeley.edu,* [b]*aditya@stat.berkeley.edu*
[2]*Google LLC,* [c]*blfang@berkeley.edu*

Multivariate adaptive regression splines (MARS) is a popular method for nonparametric regression introduced by Friedman in 1991. MARS fits simple nonlinear and non-additive functions to regression data. We propose and study a natural lasso variant of the MARS method. Our method is based on least squares estimation over a convex class of functions obtained by considering infinite-dimensional linear combinations of functions in the MARS basis and imposing a variation based complexity constraint. Our estimator can be computed via finite-dimensional convex optimization, although it is defined as a solution to an infinite-dimensional optimization problem. Under a few standard design assumptions, we prove that our estimator achieves a rate of convergence that depends only logarithmically on dimension and thus avoids the usual curse of dimensionality to some extent. We also show that our method is naturally connected to nonparametric estimation techniques based on smoothness constraints. We implement our method with a cross-validation scheme for the selection of the involved tuning parameter and compare it to the usual MARS method in various simulation and real data settings.

1. Introduction. We study a natural lasso variant of the multivariate adaptive regression splines (MARS) method (see Friedman (1991) or Hastie, Tibshirani and Friedman (2009), Section 9.4) for nonparametric regression. To understand the relationship between a response variable y and d explanatory variables $x_1, \ldots, x_d$ based on observed data $(x^{(1)}, y_1), \ldots, (x^{(n)}, y_n)$ with $x^{(i)} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, MARS fits a function $y = \hat{f}_{\text{mars}}(x_1, \ldots, x_d)$ where $\hat{f}_{\text{mars}}$ is a sparse linear combination of functions of the form

$$\prod_{j=1}^{d} (b_j(x_j))^{\alpha_j} = \prod_{j:\alpha_j=1} b_j(x_j) \tag{1}$$

with $\alpha = (\alpha_1, \ldots, \alpha_d) \in \{0, 1\}^d$ and

$$b_j(x_j) = (x_j - t_j)_+ \quad \text{or} \quad (t_j - x_j)_+ \quad \text{for some real number } t_j.$$

Here $\cdot_+ := \max\{\cdot, 0\}$ indicates the ReLU function. As a concrete example, to understand the relationship between the logarithm of Weekly Earnings (y) and the two variables, Years of Education (x_1) and Years of Experience (x_2), from a standard dataset (`ex1029` in the R library `Sleuth3`) collected from 25,437 full-time male workers in 1987, the default implementation of MARS from the R package `earth` (with the maximum degree of interaction set to two) fits

Received June 2023; revised March 2024.
MSC2020 subject classifications. 62G08.
Key words and phrases. Bracketing entropy bounds, constrained least squares estimation, curse of dimensionality, Hardy–Krause variation, infinite-dimensional optimization, integrated Brownian sheet, locally adaptive regression spline, L1 penalty, metric entropy bounds, mixed derivatives, nonparametric regression, piecewise linear function estimation, small ball probability, tensor products, total variation regularization, trend filtering.

the function

$$
\begin{aligned}
&5.83 + 0.0695(x_1 - 7)_+ - 0.0370(11 - x_1)_+ \\
&\quad + 0.0155(x_2 - 13)_+ - 0.0600(13 - x_2)_+ - 0.0164(x_2 - 30)_+ \\
&\quad - 0.0114(x_1 - 11)_+(x_2 - 40)_+ + 0.00148(x_1 - 11)_+(40 - x_2)_+,
\end{aligned} \tag{2}
$$

which is clearly a linear combination of the eight functions each of the form (1).

MARS fits a nonlinear function to the observed data that is simple enough to be interpretable because it is built from the basic ReLU functions $(x_j - t_j)_+$ and $(t_j - x_j)_+$. Furthermore, MARS fits non-additive functions because of the presence of products in (1), which enables interactions between the explanatory variables $x_1, \ldots, x_d$. Indeed, the term (1) can be interpreted as an interaction term of order $|\alpha|$ between the variables in the set $S(\alpha)$. Here and in the rest of the paper, we use the notation

$$
S(\alpha) := \{j \in [d] : \alpha_j = 1\} \quad \text{where } [d] := \{1, \ldots, d\},
$$

and

$$
|\alpha| := |S(\alpha)| = \sum_{j=1}^{d} \mathbf{1}\{\alpha_j = 1\}.
$$

The exact methodology that MARS uses involves a greedy algorithm similar to stepwise regression methods. Specifically, one adds in basis functions of the form (1) starting with a constant function using a goodness of fit criterion. Typically, one only considers terms (1) for which the interaction order $|\alpha|$ is smaller than a pre-chosen integer $s \le d$ (most commonly $s = 1$ or $s = 2$). Once a reasonably large number of basis functions with $|\alpha| \le s$ are added, a backward deletion procedure is applied to remove superfluous basis functions. We refer the reader to Hastie, Tibshirani and Friedman (2009), Section 9.4 for more details on MARS.

Our goal in this paper is to propose and study a lasso variant of the MARS method where we consider all the basis functions of the form (1) with $|\alpha| \le s$ and apply the lasso method of Tibshirani (1996). As is well known, lasso is an attractive alternative to stepwise regression methods in usual linear models. In order to apply lasso in the MARS setting, we first assume that the explanatory variables $x_1, \ldots, x_d$ all take values in the interval $[0, 1]$. In other words, the domain of the regression function is assumed to be $[0, 1]^d$. In practical settings, this can be achieved by subtracting the minimum possible value and dividing by the range for each explanatory variable. This scaling puts all the variables on a comparable footing enabling the application of lasso. Without such a scaling, $x_1, \ldots, x_d$ might be on very different scales in which case penalizing or constraining the sum of the absolute values of the coefficients corresponding to the terms $(x_j - t_j)_+$ would be unnatural (e.g., think of the setting where x_1 is *years* of education and x_2 is *days* of experience). After fitting a function in the transformed domain $[0, 1]^d$, we can simply invert the transformation to find the equation of the fitted function in the original domain (see Section 6 for some real data applications).

In the rest of the paper, we assume that the explanatory variables $x_1, \ldots, x_d$ all belong to $[0, 1]$. The observed data is $(x^{(1)}, y_1), \ldots, (x^{(n)}, y_n)$ where $x^{(i)} \in [0, 1]^d$ and $y_i \in \mathbb{R}$. To this data, we fit functions of the form $y = \hat{f}(x_1, \ldots, x_d)$ where $\hat{f} : [0, 1]^d \to \mathbb{R}$ via the application of lasso with the MARS basis functions. The restriction $x_j \in [0, 1]$ allows us to make two simplifications to the usual MARS setup:

(i) Instead of considering both kinds of functions $(x_j - t_j)_+$ and $(t_j - x_j)_+$, we only take into account $(x_j - t_j)_+$, because as each x_j is assumed to be in $[0, 1]$, we can write

$$
(t_j - x_j)_+ = (x_j - t_j)_+ - x_j + t_j = (x_j - t_j)_+ - (x_j - 0)_+ + t_j,
$$

which implies that every linear combination of functions of the form (1) is also a linear combination of functions of the same form (1) where $b_j(x_j) = (x_j - t_j)_+$ for some t_j.

(ii) We assume that $t_j \in [0,1)$ for each j. This is because when $t_j \geq 1$, the function $(x_j - t_j)_+$ becomes 0 as $x_j \in [0,1]$, and for $t_j < 0$, the function $(x_j - t_j)_+ = x_j - t_j = (x_j - 0)_+ - t_j$ is a linear combination of $(x_j - 0)_+$ and the constant function 1.

Because there are an uncountable number of functions of the form (1) (as t_j can be any number in $[0,1)$), we work with an infinite-dimensional version of lasso. Infinite-dimensional lasso formulations have been used in many papers including Rosset et al. (2007), de Castro and Gamboa (2012), Bredies and Pikkarainen (2013), Candès and Fernandez-Granda (2014), Duval and Peyré (2015), De Castro et al. (2017), Denoyelle et al. (2020), and Condat (2020), which studied various inverse problems in spaces of measures. The main idea is to consider infinite linear combinations of basis functions that are parametrized by signed measures and to measure complexity in terms of the variations of the involved signed measures. In the MARS context, infinite linear combinations of the basis functions (1) with $|\alpha| \leq s$ are

$$f_{a_{\mathbf{0}},\{\nu_\alpha\}}(x_1,\ldots,x_d) := a_{\mathbf{0}} + \sum_{\substack{\alpha\in\{0,1\}^d\setminus\{\mathbf{0}\}\\ |\alpha|\leq s}} \int_{[0,1)^{|\alpha|}} \prod_{j\in S(\alpha)} (x_j - t_j)_+ \, d\nu_\alpha\big(t^{(\alpha)}\big), \tag{3}$$

where $\mathbf{0} := (0,\ldots,0)$, $a_{\mathbf{0}} \in \mathbb{R}$, ν_α is a finite (Borel) signed measure on $[0,1)^{|\alpha|}$, and $t^{(\alpha)}$ indicates the vector $(t_j, j \in S(\alpha))$ for each binary vector $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$ with $|\alpha| \leq s$. We will denote the collection of all such functions $f_{a_{\mathbf{0}},\{\nu_\alpha\}}$ by $\mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ (the subscript ∞ highlights the fact that $\mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ contains *infinite* linear combinations of the functions (1)). The usual MARS functions are special cases of (3) corresponding to discrete signed measures ν_α. Indeed, when each ν_α is supported on a finite set $\{(t^{(\alpha)}_{lj}, j \in S(\alpha)) : l = 1,\ldots,k_\alpha\}$ with

$$\nu_\alpha\big(\big\{\big(t^{(\alpha)}_{lj}, j \in S(\alpha)\big)\big\}\big) = b^{(\alpha)}_l \quad \text{for } l = 1,\ldots,k_\alpha, \tag{4}$$

the function $f_{a_{\mathbf{0}},\{\nu_\alpha\}}$ becomes

$$(x_1,\ldots,x_d) \mapsto a_{\mathbf{0}} + \sum_{\substack{\alpha\in\{0,1\}^d\setminus\{\mathbf{0}\}\\ |\alpha|\leq s}} \sum_{l=1}^{k_\alpha} b^{(\alpha)}_l \prod_{j\in S(\alpha)} \big(x_j - t^{(\alpha)}_{lj}\big)_+. \tag{5}$$

Our infinite-dimensional lasso estimator minimizes the least squares criterion over $f_{a_{\mathbf{0}},\{\nu_\alpha\}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ with a constraint on the complexity of $f_{a_{\mathbf{0}},\{\nu_\alpha\}}$. The complexity measure involves the sum of the variations of the underlying signed measures ν_α and is an infinite-dimensional analogue of the usual L^1 norm of the coefficients used in finite-dimensional lasso. Recall that, for a signed measure ν on Ω and a measurable subset $E \subseteq \Omega$, the variation of ν on E is denoted by $|\nu|(E)$ and is defined as the supremum of $\sum_{A\in\pi} |\nu(A)|$ over all partitions π of E into a countable number of disjoint measurable subsets. Using the variation of the involved signed measures, we define our complexity measure for functions $f = f_{a_{\mathbf{0}},\{\nu_\alpha\}} \in \mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ by

$$V_{\mathrm{mars}}(f_{a_{\mathbf{0}},\{\nu_\alpha\}}) = \sum_{\substack{\alpha\in\{0,1\}^d\setminus\{\mathbf{0}\}\\ |\alpha|\leq s}} |\nu_\alpha|\big([0,1)^{|\alpha|}\setminus\{\mathbf{0}\}\big). \tag{6}$$

We are excluding $\mathbf{0} = (0,\ldots,0)$ in the variation of ν_α because we want to only penalize those basis functions that include at least one nonlinear term and leave unpenalized basis functions that are products of linear functions (note that $(x_j - t_j)_+ = x_j$ is linear when $t_j = 0$ because $x_j \in [0,1]$). In Section 11.1 of the Supplementary Material (Ki, Fang and Guntuboyina (2024)), we show that this complexity measure is well defined by proving the uniqueness of the representation $f = f_{a_{\mathbf{0}},\{\nu_\alpha\}}$.

To see why (6) is a generalization (to infinite linear combinations) of the notion of L^1 norm over the coefficients of the finite linear combination (5), just note that when ν_α is the discrete signed measure given by (4), we have

$$V_{\mathrm{mars}}(f_{a_{\mathbf{0}},\{\nu_\alpha\}}) = \sum_{\substack{\alpha\in\{0,1\}^d\setminus\{\mathbf{0}\}\\ |\alpha|\le s}} \sum_{l=1}^{k_\alpha} |b_l^{(\alpha)}| \cdot \mathbf{1}\{(t_{lj}^{(\alpha)}, j \in S(\alpha)) \neq \mathbf{0}\},$$

which is simply the sum of the absolute values of the coefficients in (5) corresponding to the basis functions that have at least one nonlinear term in their product.

Our estimator is thus given by

$$\hat{f}_{n,V}^{d,s} \in \operatorname*{argmin}_f \left\{ \sum_{i=1}^n \big(y_i - f\big(x^{(i)}\big)\big)^2 : f \in \mathcal{F}_{\infty-\mathrm{mars}}^{d,s} \text{ and } V_{\mathrm{mars}}(f) \le V \right\} \tag{7}$$

for a tuning parameter $V > 0$. We prove that $\hat{f}_{n,V}^{d,s}$ exists and can be computed by applying finite-dimensional lasso algorithms to the finite basis of functions obtained by placing the following restrictions on the knots t_j in (1):

$$t_j \in \{0\} \cup \big\{x_j^{(i)} : i \in [n]\big\}. \tag{8}$$

Here we use the notation $x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})$ for the ith design point $x^{(i)}$. As the finite-dimensional lasso estimation procedure usually zeros out many regression coefficients, it enables us to obtain $\hat{f}_{n,V}^{d,s}$ that is a sparse linear combination of (1). Therefore, our estimation procedure can be seen as an alternative to the usual MARS procedure. It is interesting to note that the usual MARS algorithm also works with the restriction (8) on the knots, although typically no theoretical justification is provided for this reduction. We also introduce a computationally more efficient approximate version $\tilde{f}_{n,V}^{d,s}$ of $\hat{f}_{n,V}^{d,s}$ which seems to work nearly as well in practice. The approximate version $\tilde{f}_{n,V}^{d,s}$ is obtained by restricting the knots t_j as

$$t_j \in \left\{0, \frac{1}{N_j}, \frac{2}{N_j}, \ldots, 1\right\}$$

for some pre-selected positive integers $N_1, \ldots, N_d$. For large n, $\tilde{f}_{n,V}^{d,s}$ can be computed much more efficiently than $\hat{f}_{n,V}^{d,s}$.

We study the theoretical accuracy of these estimators for an unknown regression function f^* under the standard regression model:

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

where ξ_i are mean zero errors whose distributions satisfy certain restrictions. We work with both the fixed design setting where $x^{(1)}, \ldots, x^{(n)}$ form a lattice in $[0,1]^d$, as well as the random design setting where $x^{(1)}, \ldots, x^{(n)}$ are assumed to be realizations of i.i.d. random variables. In the former lattice design setting, which is restrictive but standard in nonparametric function estimation (see, e.g., Nemirovski (2000)), we analyze the non-asymptotic accuracy of $\hat{f}_{n,V}^{d,s}$ and $\tilde{f}_{n,V}^{d,s}$. In the latter random design setting, we study their accuracy asymptotically.

Our theoretical results show that these estimators achieve rates of convergence of the form $n^{-4/5}(\log n)^{as+b}$ for some constants a and b. It is already known that in the univariate case ($d = s = 1$), the estimator $\hat{f}_{n,V}^{1,1}$ achieves the rate $n^{-4/5}$ (see, e.g., Mammen and van de Geer (1997), Theorem 10 or Guntuboyina et al. (2020), Theorem 2.1). Thus, our results imply that in going from the univariate to the multivariate setting, the rate of convergence only deteriorates by a logarithmic multiplicative factor. This suggests that our lasso method for

MARS fitting avoids the usual curse of dimensionality to some extent and can thus be an effective function estimation technique in higher dimensions.

We can see why our estimators achieve the dimension-free rates (up to the logarithmic multiplicative factors) in part from an alternative characterization of $\hat{f}^{d,s}_{n,V}$. We can characterize $\hat{f}^{d,s}_{n,V}$ alternatively as a least squares estimator over a class of functions whose smoothness order, in a certain sense, grows with the dimension d. A key role in this characterization is played by mixed partial derivatives of order 2. For an integer $k \geq 1$ and a real-valued function f defined on $[0,1]^m$, by the mixed partial derivatives of f of order k, we mean

$$f^{(\beta)} := \frac{\partial^{\beta_1+\cdots+\beta_m} f}{\partial x_1^{\beta_1} \cdots \partial x_m^{\beta_m}}, \tag{9}$$

where β is an m-dimensional nonnegative integer vector with $\max_j \beta_j = k$. Whenever we use the notation $f^{(\beta)}$, we inherently assume that f is sufficiently smooth, so that the right-hand side of (9) is irrespective of the order of differentiation and $f^{(\beta)}$ is well defined. Using mixed partial derivatives, we prove the following alternative characterization of $\hat{f}^{d,s}_{n,V}$:

$$\begin{aligned}\hat{f}^{d,s}_{n,V} \in \operatorname*{argmin}_f \Bigg\{ &\sum_{i=1}^n \big(y_i - f\big(x^{(i)}\big)\big)^2 : \sum_{\substack{\beta \in \{0,1,2\}^d \\ \max_j \beta_j = 2}} \int_{\bar{T}^{(\beta)}} |f^{(\beta)}| \leq V \\ &\text{and } f^{(\alpha)} = 0 \text{ for every } \alpha \in \{0,1\}^d \text{ with } |\alpha| > s \Bigg\},\end{aligned} \tag{10}$$

where

$$\bar{T}^{(\beta)} := \bar{T}_1^{(\beta)} \times \cdots \times \bar{T}_d^{(\beta)} \quad \text{where } \bar{T}_k^{(\beta)} = \begin{cases} [0,1] & \text{if } \beta_k = \max_j \beta_j \\ \{0\} & \text{otherwise.} \end{cases} \tag{11}$$

The main condition here is that the sum of the L^1 norms of mixed partial derivatives of order 2 is at most V. The set $\bar{T}^{(\beta)}$ appearing in the integral signifies that the integral of the mixed partial derivative $f^{(\beta)}$ is only over those coordinates x_l for which $\beta_l = \max_j \beta_j$ (the remaining coordinates are set to zero). Also, the condition $f^{(\alpha)} = 0$ for $|\alpha| > s$ rules out interactions of order greater than s. This characterization shows that the maximum total order $\beta_1 + \cdots + \beta_d$ of the mixed partial derivatives appearing in the constraint equals $2d$. In this sense, the smoothness order of the constraint can be taken to be $2d$, which increases with the dimension d and explains the dimension-free (up to the logarithmic multiplicative factors) rates of convergence. It should be noted however that not all (in fact, only one) mixed partial derivatives of total order $2d$ are considered in the constraint, and this keeps the function class being too small or restrictive. Also, it should be mentioned that it is well known from approximation theory that L^p norm constraints on mixed partial derivatives are advantageous and allow one to overcome the curse of dimensionality to some extent from the perspective of metric entropy, approximation, and interpolation (see, e.g., Bungartz and Griebel (2004), Dũng, Temlyakov and Ullrich (2018), Temlyakov (2018)).

In fact, the smoothness characterization (10) is not fully rigorous. Functions of the form (2) clearly belong to the constraint set in (7), but they do not belong to the constraint set in (10) because mixed partial derivatives of order 2 do not exist for these functions. We fix this problem by interpreting the L^1 norms of mixed partial derivatives of order 2 in terms of the Hardy–Krause variations of particular derivatives that we will define in Section 4. Hardy–Krause variation (see, e.g., Aistleitner and Dick (2015), Owen (2005)) is a multivariate generalization of total variation of univariate functions (we review the definition of

Hardy–Krause variation and its properties in Section 7 of the Supplementary Material (Ki, Fang and Guntuboyina (2024))). Thus, even though (10) is not fully rigorous because mixed partial derivatives of order 2 do not exist for many important MARS functions, it is still helpful for understanding how the curse of dimensionality can be avoided by our estimators.

The characterization (10) also connects our estimators to other related methods from the literature. In the univariate case ($d = s = 1$), we have

$$\hat{f}^{1,1}_{n,V} \in \operatorname*{argmin}_f \left\{ \sum_{i=1}^{n} \big(y_i - f\big(x^{(i)}\big)\big)^2 : \int_0^1 |f''| \le V \right\}$$

which is a constrained analogue of the locally adaptive regression splines estimator of Mammen and van de Geer (1997) when the order k (in their notation) equals 2. Hence, our estimator $\hat{f}^{d,s}_{n,V}$ can be seen as a multivariate generalization of this univariate estimator of Mammen and van de Geer (1997). Furthermore, if $s = d$ and the condition $\max_j \beta_j = 2$ is replaced with $\max_j \beta_j = 1$ in (10), then one obtains the Hardy–Krause variation denoising estimator of Fang, Guntuboyina and Sen (2021). Therefore, we can also view $\hat{f}^{d,d}_{n,V}$ as a second-order Hardy–Krause variation denoising estimator. Further connections to related work are detailed in Section 5.

We would like to point out here that the theoretical rates of convergence as well as the smoothness characterization have been made possible due to our infinite-dimensional lasso formulation of MARS. In contrast, to the best of our knowledge, no rates of convergence are known for the usual MARS method. Also, there exist no prior connections between the usual MARS method and nonparametric regression methods based on smoothness assumptions.

In addition to the above theoretical contributions, we also implement our method with a cross-validation scheme for the selection of the tuning parameter V and compare our estimators to the usual MARS estimator using simulated and real data.

The rest of the paper is organized as follows. In Section 2, we present results on the existence and computation of $\hat{f}^{d,s}_{n,V}$ and also introduce the approximate version $\tilde{f}^{d,s}_{n,V}$. Theoretical accuracy results for $\hat{f}^{d,s}_{n,V}$ and $\tilde{f}^{d,s}_{n,V}$ are in Section 3. Section 4 is devoted to the alternative characterization (10) based on smoothness. In Section 5, we discuss connections between our method and other related methods. In Section 6, we illustrate the performance of our method in simulated and real data settings and compare its performance to that of the usual MARS algorithm.

2. Existence, computation, and approximation. In this section, we prove the existence of our infinite-dimensional lasso estimator $\hat{f}^{d,s}_{n,V}$ (defined in (7)) and show that it can be computed via finite-dimensional lasso 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 (7) only depends on the function f through its values at the design points $x^{(i)}$, $i \in [n]$. 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) where each ν_α is a discrete signed measure supported on the lattice generated by the design points. For each $k \in [d]$, let $\mathcal{U}_k$ denote the finite subset of $[0, 1]$ consisting of the points $0, x^{(1)}_k, \ldots, x^{(n)}_k, 1$ (recall here that $x^{(i)}_k$ denotes the kth coordinate of the ith design point $x^{(i)} = (x^{(i)}_1, \ldots, x^{(i)}_d)$). As there could be ties among $0, x^{(1)}_k, \ldots, x^{(n)}_k, 1$, we will write, for some $n_k \in [n+1]$,

$$\mathcal{U}_k = \{u^{(k)}_0, u^{(k)}_1, \ldots, u^{(k)}_{n_k}\} \quad \text{where } 0 = u^{(k)}_0 < \cdots < u^{(k)}_{n_k} = 1.$$

Note specially that the cardinality of $\mathcal{U}_k$ is $n_k + 1$, that $u^{(k)}_0$ is always 0, and that $u^{(k)}_{n_k}$ is always 1. The next lemma (proved in Section 11.2.1 of the Supplementary Material (Ki, Fang and

Guntuboyina (2024))) implies that, for the optimization problem (7), we can restrict to the functions of the form (3) where each ν_α is a discrete signed measure supported on the finite set $(\prod_{k\in S(\alpha)} \mathcal{U}_k) \cap [0,1)^{|\alpha|}$.

LEMMA 2.1. *Suppose we are given a real number $a_{\mathbf{0}}$ and a collection of finite signed measures $\{\nu_\alpha\}$ where ν_α is defined on $[0,1)^{|\alpha|}$ for each $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$ with $|\alpha| \le s$. Then, there exists a collection of discrete signed measures $\{\mu_\alpha\}$ where μ_α is concentrated on $(\prod_{k\in S(\alpha)} \mathcal{U}_k) \cap [0,1)^{|\alpha|}$ for each $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$ with $|\alpha| \le s$ such that*

(i) $f_{a_{\mathbf{0}},\{\mu_\alpha\}}(x^{(i)}) = f_{a_{\mathbf{0}},\{\nu_\alpha\}}(x^{(i)})$ *for all* $i \in [n]$, *and*
(ii) $V_{\mathrm{mars}}(f_{a_{\mathbf{0}},\{\mu_\alpha\}}) \le V_{mars}(f_{a_{\mathbf{0}},\{\nu_\alpha\}})$.

When ν_α is concentrated on $(\prod_{k\in S(\alpha)} \mathcal{U}_k) \cap [0,1)^{|\alpha|}$ for each α, the function $f_{a_{\mathbf{0}},\{\nu_\alpha\}}$ can be written as

$$a_{\mathbf{0}} + \sum_{\substack{\alpha\in\{0,1\}^d\setminus\{\mathbf{0}\}\\ |\alpha|\le s}} \sum_{l\in\prod_{k\in S(\alpha)}[0:(n_k-1)]} \nu_\alpha\big(\{(u_{l_k}^{(k)}, k\in S(\alpha))\}\big) \cdot \prod_{k\in S(\alpha)} (x_k - u_{l_k}^{(k)})_+, \tag{12}$$

where we use the notation $[p:q] := \{p, p+1, \dots, q\}$ for two integers $p \le q$. Also, its complexity measure becomes

$$V_{\mathrm{mars}}(f_{a_{\mathbf{0}},\{\nu_\alpha\}}) = \sum_{\substack{\alpha\in\{0,1\}^d\setminus\{\mathbf{0}\}\\ |\alpha|\le s}} \sum_{l\in\prod_{k\in S(\alpha)}[0:(n_k-1)]\setminus\{\mathbf{0}\}} \big|\nu_\alpha\big(\{(u_{l_k}^{(k)}, k\in S(\alpha))\}\big)\big|.$$

The above function (12) is a linear combination of the basis functions (1) whose knots t_k are chosen from $\mathcal{U}_k \setminus \{1\} = \{0, x_k^{(1)}, \dots, x_k^{(n)}\}$, and its complexity measure equals the absolute sum of the coefficients of the involved basis functions with at least one nonlinear term. Thus, if we additionally assume that f in the problem (7) is constructed from discrete signed measures as above, then (7) reduces to a finite-dimensional lasso problem. Lemma 2.1 then implies that every solution to this finite-dimensional lasso problem is also a solution to (7). A precise statement is given in the following result, which we prove in Section 11.2.2 of the Supplementary Material (Ki, Fang and Guntuboyina (2024)).

PROPOSITION 2.2. *Let*

$$J = \Big\{(\alpha, l) : \alpha \in \{0,1\}^d \setminus \{\mathbf{0}\},\ |\alpha| \le s,\ \text{and } l \in \prod_{k\in S(\alpha)} [0:(n_k-1)]\Big\}$$

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

$$M_{i,(\alpha,l)} = \prod_{k\in S(\alpha)} (x_k^{(i)} - u_{l_k}^{(k)})_+ \quad \text{for } i \in [n] \text{ and } (\alpha, l) \in J.$$

Also, let $(\hat{a}_{\mathbf{0}}, \hat{\gamma}_{n,V}^{d,s}) \in \mathbb{R} \times \mathbb{R}^{|J|}$ be a solution to the following finite-dimensional lasso problem

$$(\hat{a}_{\mathbf{0}}, \hat{\gamma}_{n,V}^{d,s}) \in \operatorname*{argmin}_{a_{\mathbf{0}}\in\mathbb{R}, \gamma\in\mathbb{R}^{|J|}} \Big\{\|y - a_{\mathbf{0}}\mathbf{1} - M\gamma\|_2^2 : \sum_{\substack{(\alpha,l)\in J\\ l\ne\mathbf{0}}} |\gamma_{\alpha,l}| \le V\Big\}, \tag{13}$$

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

$$f(x_1, \dots, x_d) = \hat{a}_{\mathbf{0}} + \sum_{(\alpha,l)\in J} (\hat{\gamma}_{n,V}^{d,s})_{\alpha,l} \cdot \prod_{k\in S(\alpha)} (x_k - u_{l_k}^{(k)})_+ \tag{14}$$

is a solution to the problem (7). *The problem* (7) *can have multiple solutions, but every solution* $\hat{f}^{d,s}_{n,V}$ *satisfies*

$$\hat{f}^{d,s}_{n,V}(x^{(i)}) = \hat{a}_{\mathbf{0}} + (M\hat{\gamma}^{d,s}_{n,V})_i = \hat{a}_{\mathbf{0}} + \sum_{(\alpha,l)\in J} (\hat{\gamma}^{d,s}_{n,V})_{\alpha,l} \cdot \prod_{k\in S(\alpha)} (x^{(i)}_k - u^{(k)}_{l_k})_+$$

for every $i \in [n]$.

Because the set

$$\left\{ a_0\mathbf{1} + M\gamma : a_0 \in \mathbb{R}, \gamma \in \mathbb{R}^{|J|}, \text{ and } \sum_{\substack{(\alpha,l)\in J \\ l\neq\mathbf{0}}} |\gamma_{\alpha,l}| \le V \right\}$$

is closed and convex, there exists a solution to the finite-dimensional lasso problem (13). Hence, the existence of solutions to our estimation problem (7) is guaranteed by Proposition 2.2. Also, once we find a solution to the problem (13) via any optimization algorithms, we can construct a solution $\hat{f}^{d,s}_{n,V}$ to the problem (7) through the equation (14).

However, solving the finite-dimensional lasso problem (13) can be computationally intensive if n is large because the number of columns of M equals

$$|J| = \sum_{\substack{\alpha\in\{0,1\}^d\setminus\{\mathbf{0}\} \\ |\alpha|\le s}} \prod_{k\in S(\alpha)} n_k,$$

which is of order $O(n^s)$ (ignoring a multiplicative factor in d) in the worst case when each $n_k = O(n)$. In the current implementation of our method, we utilize the optimization software MOSEK as a black-box tool for solving the problem (13) (see Section 6 for more details). Using this black-box tool involves creating the whole matrix 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 2.1 ensures that we only need to consider discrete signed measures ν_α supported on the lattices $(\prod_{k\in S(\alpha)} \mathcal{U}_k) \cap [0,1)^{|\alpha|}$ for our estimation problem (7). In the approximate method, we instead restrict our attention to discrete signed measures ν_α supported on the lattices generated by

$$\tilde{\mathcal{U}}_k = \left\{ 0, \frac{1}{N_k}, \frac{2}{N_k}, \ldots, 1 \right\}$$

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

We then consider the finite-dimensional optimization problem to which the problem (7) reduces when we additionally impose such restrictions on signed measures ν_α. We call this problem the approximate (finite-dimensional optimization) problem. The approximate problem has the same form as (13) but with different M and J. Here

$$J = \left\{ (\alpha, l) : \alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}, |\alpha| \le s, \text{ and } l \in \prod_{k\in S(\alpha)} [0 : (N_k - 1)] \right\},$$

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

$$M_{i,(\alpha,l)} = \prod_{k\in S(\alpha)} \left(x^{(i)}_k - \frac{l_k}{N_k} \right)_+ \quad \text{for } i \in [n] \text{ and } (\alpha, l) \in J.$$

As opposed to the original finite-dimensional problem (13), the number of columns of M in this problem is always fixed and not affected by the design points $x^{(1)}, \dots, x^{(n)}$. Hence, this approximate problem can be solved much efficiently than (13), 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 the equation (14) as before. We denote this estimator by $\tilde{f}^{d,s}_{n,V}$ and call it an approximate version of $\hat{f}^{d,s}_{n,V}$. In the next section, we study the theoretical accuracy of $\tilde{f}^{d,s}_{n,V}$ along with $\hat{f}^{d,s}_{n,V}$. We will see that if we choose N_k appropriately, the approximate method is as accurate as the original method, while it significantly improves computational efficiency.

3. Risk analysis. This section is dedicated to the study of the theoretical accuracy of $\hat{f}^{d,s}_{n,V}$ and $\tilde{f}^{d,s}_{n,V}$ as an estimator for unknown regression functions. We first consider the nonasymptotic accuracy of $\hat{f}^{d,s}_{n,V}$ and $\tilde{f}^{d,s}_{n,V}$ in the fixed design setting and then study their asymptotic accuracy in the random design setting. The proofs of all the results in this section are provided in Section 11.3 of the Supplementary Material (Ki, Fang and Guntuboyina (2024)).

3.1. *Fixed design.* Here we assume that $x^{(1)}, \dots, x^{(n)}$ form a lattice

$$\{x^{(1)}, \dots, x^{(n)}\} = \prod_{k=1}^{d} \{u^{(k)}_{i_k} : i_k \in [0 : (n_k - 1)]\}, \tag{15}$$

where for every $k \in [d]$, we have $n_k \geq 2$, $0 = u^{(k)}_0 < u^{(k)}_1 < \cdots < u^{(k)}_{n_k - 1} \leq 1$, and

$$u^{(k)}_{i_k} - u^{(k)}_{i_k - 1} \geq \frac{\rho}{n_k} \quad \text{for all } i_k \in [n_k - 1]$$

for some constant $\rho > 0$. We also assume that $y_1, \dots, y_n$ are generated according to the regression model

$$y_i = f^*(x^{(i)}) + \xi_i, \tag{16}$$

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

$$\mathbb{E}[e^{\lambda \xi_i}] \leq e^{\frac{\sigma^2 \lambda^2}{2}}$$

for all $\lambda \in \mathbb{R}$. We measure the accuracy of an estimator $\hat{f}_n$ of f^* via the squared empirical L^2 norm

$$\|\hat{f}_n - f^*\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n(x^{(i)}) - f^*(x^{(i)}))^2 \tag{17}$$

and define its risk as

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

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

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

THEOREM 3.1. *Suppose $f^* \in \mathcal{F}^{d,s}_{\infty - \text{mars}}$ and $V_{\text{mars}}(f^*) \leq V$ and assume the lattice design (15). The estimator $\hat{f}^{d,s}_{n,V}$ then satisfies that*

$$\mathcal{R}_F(\hat{f}^{d,s}_{n,V}, f^*) \leq C_{\rho,d} \left(\frac{\sigma^2 V^{\frac{1}{2}}}{n} \right)^{\frac{4}{5}} \left[\log\left(2 + \frac{V n^{\frac{1}{2}}}{\sigma} \right) \right]^{\frac{3(2s-1)}{5}} + C_{\rho,d} \frac{\sigma^2}{n} [\log n]^2 \tag{18}$$

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 (18), so that

$$\mathcal{R}_F\big(\hat{f}^{d,s}_{n,V}, f^*\big) = O\big(n^{-\frac{4}{5}}(\log n)^{\frac{3(2s-1)}{5}}\big),$$

where the multiplicative constant underlying $O(\cdot)$ depends on ρ, d, σ, and V. In the univariate case, we can deduce from Guntuboyina et al. (2020), Theorem 2.1 that

$$\mathcal{R}_F\big(\hat{f}^{1,1}_{n,V}, f^*\big) \le C_\rho\bigg(\frac{\sigma^2 V^{\frac{1}{2}}}{n}\bigg)^{\frac{4}{5}} + C_\rho\frac{\sigma^2}{n}\log n,$$

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

$$\mathcal{R}_F\big(\hat{f}^{1,1}_{n,V}, f^*\big) = O\big(n^{-\frac{4}{5}}\big),$$

where the multiplicative constant underlying $O(\cdot)$ depends on ρ, σ, and V. Thus, what Theorem 3.1 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 multiplicative factor depending on s. This suggests that our lasso method for MARS fitting 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 3.1 is to find an upper bound of the metric entropy of $\mathcal{D}_m$ (under the L^2 norm), which is defined as the collection of all the functions of the form

$$(x_1, \ldots, x_m) \mapsto \int (x_1 - t_1)_+ \cdots (x_m - t_m)_+ \, d\nu(t),$$

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

THEOREM 3.2. *There exist positive constants C_m and ϵ_m depending on m such that*

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

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

REMARK 3.3. 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 the functions of the form

$$(x_1, \ldots, x_m) \mapsto \int \mathbf{1}\{x_1 \ge t_1\} \cdots \mathbf{1}\{x_m \ge t_m\} \, d\nu(t) = \nu\big([\mathbf{0}, x]\big).$$

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 Blei, Gao and Li (2007). Thus, we are basically extending the argument in Blei, Gao and Li (2007) from $\mathbf{1}\{x \ge t\}$ to $(x-t)_+$.

Theorem 3.2 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 Blei, Gao and Li (2007), Section 3 and Gao (2008), Section 3 and results from Li and Linde (1999), Theorem 1.2 and Artstein et al. (2004), Theorem 5. The small ball probability of integrated Brownian sheet here refers to the quantity

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

where $\epsilon > 0$ and X_m is an m-dimensional integrated Brownian sheet (a description of integrated Brownian sheet is given in Section 11.3.2 of the Supplementary Material (Ki, Fang

and Guntuboyina (2024))). Required bounds on this small ball probability are then obtained using results from Dunker et al. (1999), Theorem 6 and Chen and Li (2003), Theorem 1.2. Specifically, we show that there exist positive constants c_m and ϵ_m depending on m such that

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

for every $0 < \epsilon < \epsilon_m$. This result, along with the connection between the metric entropy and the small ball probability, leads to Theorem 3.2, which is the main ingredient in our proof of Theorem 3.1.

Now, we turn to the result for the approximate version $\tilde{f}^{d,s}_{n,V}$. The next theorem presents an upper bound of the risk of $\tilde{f}^{d,s}_{n,V}$ under the same assumption as in Theorem 3.1. Recall that N_k are the pre-selected integers for the approximate method.

THEOREM 3.4. *Suppose $f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ and $V_{\mathrm{mars}}(f^*) \le V$ and assume the lattice design (15). The estimator $\tilde{f}^{d,s}_{n,V}$ then satisfies that*

$$\mathcal{R}_F\big(\tilde{f}^{d,s}_{n,V}, f^*\big) \le \frac{8V^2}{N^2} + C_{\rho,d}\left(\frac{\sigma^2 V^{\frac{1}{2}}}{n}\right)^{\frac{4}{5}} \left[\log\left(2 + \frac{V n^{\frac{1}{2}}}{\sigma}\right)\right]^{\frac{3(2s-1)}{5}} + C_{\rho,d}\frac{\sigma^2}{n}[\log n]^2$$

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

Theorem 3.4 shows that $\tilde{f}^{d,s}_{n,V}$ 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, $\tilde{f}^{d,s}_{n,V}$ achieves the same rate as $\hat{f}^{d,s}_{n,V}$. Indeed, if we set each N_k to be of order at least $n^{2/5}$, then

$$\mathcal{R}_F\big(\tilde{f}^{d,s}_{n,V}, f^*\big) = O\big(n^{-\frac{4}{5}} (\log n)^{\frac{3(2s-1)}{5}}\big), \tag{19}$$

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

3.2. *Random design.* Here we assume that $x^{(1)}, \dots, x^{(n)}$ are realizations of i.i.d. random variables $X^{(1)}, \dots, X^{(n)}$ with a probability density function p_0 on $[0,1]^d$ that is bounded by some constant $B \ge 1$, that is, $\|p_0\|_\infty \le B$. Also, we assume that $(X^{(1)}, y_1), \dots, (X^{(n)}, y_n)$ are generated according to the regression model

$$y_i = f^*\big(X^{(i)}\big) + \xi_i, \tag{20}$$

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

$$\|\xi_i\|_{5,1} := \int_0^\infty \big(\mathbb{P}(|\xi_i| > t)\big)^{\frac{1}{5}}\, dt < \infty. \tag{21}$$

Note that the condition (21) 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., Ledoux and Talagrand (1991), Chapter 10). In this setting, we measure the accuracy of an estimator $\hat{f}_n$ of f^* by

$$\big\|\hat{f}_n - f^*\big\|^2_{p_0,2} := \int \big(\hat{f}_n(x) - f^*(x)\big)^2 p_0(x)\, dx. \tag{22}$$

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

THEOREM 3.5. *If $f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ and $V_{\mathrm{mars}}(f^*) \le V$, then we have*

$$\|\hat{f}^{d,s}_{n,V} - f^*\|^2_{p_0,2} = O_p\big(n^{-\frac{4}{5}}(\log n)^{\frac{8(s-1)}{5}}\big). \tag{23}$$

As the metric entropy of $\mathcal{D}_m$ played a central role in our proof of Theorem 3.1, the bracketing entropy of $\mathcal{D}_m$ is the key ingredient of our proof of Theorem 3.5. The following theorem states an upper bound of the bracketing entropy of $\mathcal{D}_m$.

THEOREM 3.6. *There exists a positive constant C_m depending on m such that*

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

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 3.7. Theorem 3.6 also provides an upper bound of the metric entropy of $\mathcal{D}_m$ (under the L^2 norm). Since

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

we can derive from Theorem 3.6 that

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

for every $\epsilon > 0$. However, this upper bound is weaker than the one we achieved in Theorem 3.2. Although it has the same order for ϵ, the exponent of the logarithmic multiplicative factor is bigger. We can obtain from this result an upper bound of the risk of $\hat{f}^{d,s}_{n,V}$ under the fixed lattice design setting, but it will lead to a bound looser than the one in Theorem 3.1.

We can prove a similar result as in Theorem 3.5 for the approximate version $\tilde{f}^{d,s}_{n,V}$. As we state in the following theorem, $\tilde{f}^{d,s}_{n,V}$ achieves the same rate of convergence as $\hat{f}^{d,s}_{n,V}$ if $N_1, \ldots, N_d$ are sufficiently large. Together with (19), this result suggests that the approximate method with appropriately chosen $N_1, \ldots, N_d$ can be as accurate as the original method.

THEOREM 3.8. *Suppose $f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ and $V_{\mathrm{mars}}(f^*) \le V$. Also, assume that $N = \min_k N_k = \Omega(n^{4/15})$, that is, there exists a positive constant $c_{B,d,V}$ possibly depending on B, d, and V such that*

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

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

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

Our next result shows that the logarithmic multiplicative factor in (23) can not be completely removed in the minimax sense. Specifically, we bound the minimax risk defined as

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

where the expectation is taken over $(X^{(1)}, y_1), \ldots, (X^{(n)}, y_n)$ of (20) and $\inf_{\hat{f}_n}$ denotes the infimum over all estimators $\hat{f}_n$ of f^* based on $(X^{(1)}, y_1), \ldots, (X^{(n)}, y_n)$. Here we further

restrict that ξ_i in the model (20) are independent Gaussian errors with mean zero and variance σ^2 and that the probability density function p_0 of $X^{(i)}$ is bounded below by some positive constant b, that is, $\|p_0\|_\infty \geq b$. Our result shows that the supremum risk of every estimator indeed requires a logarithmic multiplicative 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 3.9. *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*

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

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

Our proof of Theorem 3.9 is based on Assouad's lemma with a finite set of functions in $\{f^* \in \mathcal{F}^{d,s}_{\infty-\mathrm{mars}} : V_{\mathrm{mars}}(f^*) \leq V\}$ that is constructed by an extension of the ideas in Blei, Gao and Li (2007), Section 4. Results similar to Theorem 3.9 can be proved under the fixed design setting, but we do not go into detail in this paper.

4. Characterization in terms of smoothness. In this section, we provide alternative characterizations of $\mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$, $V_{\mathrm{mars}}(\cdot)$, 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 $d = s = 1$. We include the proofs of all the results in this section in Section 11.4 of the Supplementary Material (Ki, Fang and Guntuboyina (2024)).

4.1. *Smoothness characterization for $d = s = 1$.* For $d = s = 1$, $\mathcal{F}^{1,1}_{\infty-\mathrm{mars}}$ consists of all the functions $f : [0,1] \to \mathbb{R}$ of the form

$$f(x) = a_0 + \int_{[0,1)} (x-t)_+ \, d\nu(t), \tag{24}$$

where a_0 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, $V_{\mathrm{mars}}(f) = |\nu|((0,1))$.

The following simple arguments show that $\mathcal{F}^{1,1}_{\infty-\mathrm{mars}}$ can be characterized in terms of smoothness. First, by replacing $(x-t)_+$ with $\int_0^1 \mathbf{1}\{t \leq s \leq x\}\, ds$ in the integral in (24) and changing the order of integration, we obtain

$$f(x) = a_0 + \int_0^x g(t)\, dt, \tag{25}$$

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

$$g(t) = \nu\big([0,t] \cap [0,1)\big). \tag{26}$$

It can be readily verified that the function g in (26) is right-continuous on $[0,1]$ and left-continuous at 1, and the total variation $V(g)$ of g is finite and can be represented as

$$V(g) = |\nu|\big((0,1)\big). \tag{27}$$

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

$$V(h) = \sup_{0=u_0<u_1<\cdots<u_k=1} \sum_{i=0}^{k-1} \big|h(u_{i+1}) - h(u_i)\big|,$$

where the supremum is over all integers $k \geq 1$ and partitions $0 = u_0 < u_1 < \cdots < u_k = 1$ 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 (26) for a unique signed measure ν on $[0, 1)$ (see, e.g., Aistleitner and Dick (2015), Theorem 3). Putting these observations together, we can argue that $\mathcal{F}^{1,1}_{\infty-\text{mars}}$ has the following alternative characterization:

$$\mathcal{F}^{1,1}_{\infty-\text{mars}} = \Big\{ f : [0,1] \to \mathbb{R} : \exists a_0 \in \mathbb{R} \text{ and } g : [0,1] \to \mathbb{R} \text{ s.t.} \tag{28}$$
$$g \text{ is right-continuous on } [0,1), \text{ left-continuous at 1},$$
$$V(g) < \infty, \text{ and } f(x) = a_0 + \int_0^x g(t)\,dt \text{ for all } x \in [0,1] \Big\}.$$

Moreover, we can see that the complexity measure $V_{\text{mars}}(f)$ for $f \in \mathcal{F}^{1,1}_{\infty-\text{mars}}$ is equal to the total variation $V(g)$ of the function g appearing in (28).

For every function $f \in \mathcal{F}^{1,1}_{\infty-\text{mars}}$, we can show that g satisfying the conditions in (28) is unique, and thus, we can consider such g as a particular derivative of f satisfying (25). If we denote it by $D^{(1)} f$, the estimator $\hat{f}^{1,1}_{n,V}$ then can be alternatively written as

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

The representation (25) implies, by the Lebesgue differentiation theorem (see, e.g., Rudin (1987), Theorem 7.10), that f' exists and is equal to $D^{(1)} f$ 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

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

The corresponding penalized version

$$\operatorname*{argmin}_f \Big\{ \sum_{i=1}^n \big(y_i - f\big(x^{(i)}\big)\big)^2 + \lambda V\big(f'\big) \Big\} \tag{29}$$

was proposed by Mammen and van de Geer (1997) 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 Mammen and van de Geer (1997) when the order k (in their notation) equals 2. Steidl, Didas and Neumann (2006) used the terminology *second-order total variation regularization*, and Kim et al. (2009) and Tibshirani (2014) used the terminology *first-order trend filtering* for (29). Therefore, our estimator $\hat{f}^{d,s}_{n,V}$ can be considered as a multivariate generalization of piecewise linear (second-order) locally adaptive regression splines, second-order total variation regularization, or first-order trend filtering.

From the alternative characterization of $\mathcal{F}^{1,1}_{\infty-\text{mars}}$ given above, it follows that every sufficiently smooth function $f : [0, 1] \to \mathbb{R}$ belongs to $\mathcal{F}^{1,1}_{\infty-\text{mars}}$. 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

$$V(f') = \int_0^1 |f''| < \infty.$$

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

$$V_{\text{mars}}(f) = \int_0^1 |f''|. \tag{30}$$

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

4.2. *Smoothness characterization for general d and s.* As we have seen in the previous subsection, in the univariate case, $\mathcal{F}^{1,1}_{\infty-\text{mars}}$ consists of all the functions f satisfying (25) 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, which is an extension of total variation of univariate functions to higher dimensions. In Section 7 of the Supplementary Material (Ki, Fang and Guntuboyina (2024)), we review the definition of Hardy–Krause variation and its properties that we will use for proving the results in this subsection. Standard references for Hardy–Krause variation are Aistleitner and Dick (2015) and Owen (2005). Here we use Hardy–Krause variation anchored at $\mathbf{0}$, which we denote by $V_{\text{HK0}}(\cdot)$.

The following result provides an alternative characterization of $\mathcal{F}^{d,s}_{\infty-\text{mars}}$ and $V_{\text{mars}}(\cdot)$ in terms of smoothness. Recall that we use the notation $t^{(\alpha)}$ to indicate the vector $(t_j, j \in S(\alpha))$ for each $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$ with $|\alpha| \le s$.

PROPOSITION 4.1. *The function class $\mathcal{F}^{d,s}_{\infty-\text{mars}}$ consists precisely of all the functions of the form*

$$f(x_1, \ldots, x_d) = a_{\mathbf{0}} + \sum_{\substack{\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\} \\ |\alpha| \le s}} \int_{[\mathbf{0}, x^{(\alpha)}]} g_\alpha\big(t^{(\alpha)}\big)\, dt^{(\alpha)} \tag{31}$$

for some $a_{\mathbf{0}} \in \mathbb{R}$ and some collection of functions $\{g_\alpha : \alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$ and $|\alpha| \le s\}$, where for each $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$ with $|\alpha| \le s$,

(i) *g_α is a real-valued function on $[0,1]^{|\alpha|}$,*
(ii) *$V_{\text{HK0}}(g_\alpha) < \infty$,*
(iii) *g_α is coordinatewise right-continuous on $[0,1]^{|\alpha|}$, and*
(iv) *g_α is coordinatewise left-continuous at each point $x^{(\alpha)} = (x_j, j \in S(\alpha)) \in [0,1]^{|\alpha|} \setminus [0,1)^{|\alpha|}$ with respect to all the j^{th} coordinates where $x_j = 1$.*

Furthermore, the complexity of f in (31) can be written in terms of the Hardy–Krause variations of g_α as

$$V_{\text{mars}}(f) = \sum_{\substack{\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\} \\ |\alpha| \le s}} V_{HK0}(g_\alpha). \tag{32}$$

Proposition 4.1 is completely analogous to (28) for the case $d = s = 1$. Specifically, the condition (31) is analogous to the univariate condition (25). The condition $V_{\text{HK0}}(g_\alpha) < \infty$ for each $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$ with $|\alpha| \le s$ corresponds to the univariate condition $V(g) < \infty$. The coordinatewise right-continuity of each g_α on $[0,1]^{|\alpha|}$ is matched with the univariate right-continuity on $[0,1]$. Lastly, the coordinatewise left-continuity of each g_α at each $x^{(\alpha)} \in [0,1]^{|\alpha|} \setminus [0,1)^{|\alpha|}$ (with respect to all the j^{th} coordinates where $x_j = 1$) is a counterpart of the univariate 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_α over $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$ with $|\alpha| \le s$.

For every function $f \in \mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$, it can be easily checked that g_α appearing in Proposition 4.1 are uniquely determined by f. As in the case $d = s = 1$, we can thus consider such g_α as particular derivatives of f satisfying (31). Let us denote them by $D^{(\alpha)}f$ for $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$. We can then write our estimator $\hat{f}^{d,s}_{n,V}$ alternatively as

$$\hat{f}^{d,s}_{n,V} \in \operatorname*{argmin}_f \left\{ \sum_{i=1}^n \bigl(y_i - f\bigl(x^{(i)}\bigr)\bigr)^2 : \sum_{\substack{\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\} \\ |\alpha| \le s}} V_{\mathrm{HK0}}\bigl(D^{(\alpha)}f\bigr) \le V \right\}. \tag{33}$$

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 $D^{(\alpha)}$.

Recall that the univariate condition (25) implies that f' exists and equals to $D^{(1)}f$ almost everywhere. Similarly, the condition (31) imposes a certain kind of smoothness on f and characterizes the corresponding derivatives in terms of $D^{(\alpha)}f$. For each $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$ and for $x^{(\alpha)} = (x_k, k \in S(\alpha))$, let

$$D_\alpha f\bigl(x^{(\alpha)}\bigr) = \lim_{\epsilon \to 0} \frac{1}{\epsilon^{|\alpha|}} \cdot \sum_{\delta \in \prod_{k \in S(\alpha)} \{0,1\}} (-1)^{\sum_{k \in S(\alpha)} \delta_k} f\bigl(\widetilde{x(x+\epsilon)}^{(\alpha)}_\delta\bigr),$$

if the limit exists, where

$$\bigl(\widetilde{x(x+\epsilon)}^{(\alpha)}_\delta\bigr)_k = \begin{cases} \delta_k x_k + (1-\delta_k)(x_k + \epsilon) & \text{if } k \in S(\alpha) \\ 0 & \text{otherwise} \end{cases}$$

for $k \in [d]$. For example, if $d = 3$ and $\alpha = (1,1,0)$, $D_{1,1,0}f$ is defined as

$$\begin{aligned} D_{1,1,0}f(x_1,x_2) = \lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \cdot \bigl(& f(x_1+\epsilon, x_2+\epsilon, 0) - f(x_1, x_2+\epsilon, 0) \\ & - f(x_1+\epsilon, x_2, 0) + f(x_1, x_2, 0)\bigr), \end{aligned}$$

if the limit exists. Note that in contrast to mixed partial derivatives $f^{(\alpha)}$ (defined in (9)), in which partial derivatives $\partial/\partial x_j$ are taken sequentially, here all the j^{th} coordinates where $\alpha_j = 1$ are considered simultaneously. Also, note that the remaining coordinates where $\alpha_j = 0$ are set to zero for $D_\alpha f$.

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

PROPOSITION 4.2. *Suppose that the condition* (31) *holds. Then, for each* $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$, $D_\alpha f = 0$ *if* $|\alpha| > s$, *and* $D_\alpha f = D^{(\alpha)}f$ *almost everywhere* (*with respect to the Lebesgue measure*) *on* $[0,1]^{|\alpha|}$ *if* $|\alpha| \le s$.

Proposition 4.1 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 $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$, we let J_α be the set of all $\beta \in \{0,1,2\}^d$ such that

$$\max_j \beta_j = 2 \quad \text{and} \quad \beta_j = \begin{cases} 0 & \text{if } \alpha_j = 0 \\ 1 \text{ or } 2 & \text{if } \alpha_j = 1. \end{cases} \tag{34}$$

Also, recall the notation $\bar{T}^{(\beta)}$ from (11).

LEMMA 4.3. *Suppose* $f : [0,1]^d \to \mathbb{R}$ *is smooth in the sense that*

(i) $f^{(\alpha)}$ *exists and is continuous on* $[0,1]^d$ *for every* $\alpha \in \{0,1\}^d$, *and*
(ii) $f^{(\beta)}$ *exists and is continuous on* $\bar{T}^{(\alpha)}$ *for every* $\beta \in J_\alpha$, *for every* $\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\}$.

Then, $f \in \mathcal{F}^{d,d}_{\infty-\mathrm{mars}}$ *and*

$$V_{\mathrm{mars}}(f) = \sum_{\substack{\beta \in \{0,1,2\}^d \\ \max_j \beta_j = 2}} \int_{\bar{T}^{(\beta)}} |f^{(\beta)}|. \tag{35}$$

Furthermore, if $f^{(\alpha)} = 0$ *for all* $\alpha \in \{0,1\}^d$ *with* $|\alpha| > s$ *in addition, then* $f \in \mathcal{F}^{d,s}_{\infty-\mathrm{mars}}$ *and*

$$V_{\mathrm{mars}}(f) = \sum_{\substack{\alpha \in \{0,1\}^d \setminus \{\mathbf{0}\} \\ |\alpha| \le s}} \sum_{\beta \in J_\alpha} \int_{\bar{T}^{(\beta)}} |f^{(\beta)}|. \tag{36}$$

Note that the integrals on the right-hand side of (35) and (36) are only over those coordinates x_l for which $\beta_l = \max_j \beta_j$ (the remaining coordinates are set to zero).

The formula (35) is a multivariate generalization of the univariate formula (30), 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 $(\beta_1 + \cdots + \beta_d)$ up to $2d$ appear in (35). 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 the logarithmic multiplicative factors), as we observed in Section 3. However, we should also note that the maximum total order $2d$ is solely achieved by the mixed partial derivative $f^{(2,\ldots,2)}$, which 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 $V_{\mathrm{HK0}}(\cdot)$ via (35). Specifically, if the condition $\max_j \beta_j = 2$ in (35) is replaced with $\max_j \beta_j = 1$, then one obtains the formula for $V_{\mathrm{HK0}}(f)$ for sufficiently smooth functions f (see Lemma 7.4 of the Supplementary Material (Ki, Fang and Guntuboyina (2024))). Hence, we can also view $V_{\mathrm{mars}}(\cdot)$ as *second-order* Hardy–Krause variation (anchored at $\mathbf{0}$).

In Section 8 of the Supplementary Material (Ki, Fang and Guntuboyina (2024)), we describe the results of this section by specializing to the case $d = s = 2$. We encourage readers who find the results in this section too abstract to refer to Section 8 of Ki, Fang and Guntuboyina (2024) for more explicit formulae.

5. Related work. Here are some connections between our paper and existing works on nonparametric regression.

As mentioned earlier, our method can be viewed as a multivariate generalization of the piecewise linear locally adaptive regression spline estimator of Mammen and van de Geer (1997) (see also Steidl, Didas and Neumann (2006)). There are other ways of generalizing the piecewise linear locally adaptive regression splines estimator to the multivariate setting as well (see, e.g., Parhi and Nowak (2021, 2023) for some recent work). Also, we have seen that our method can be considered as a multivariate extension of first-order trend filtering (see, e.g., Kim et al. (2009), Tibshirani (2014)). Ortelli and van de Geer (2021) and Sadhanala et al. (2021) recently studied different multivariate extensions of trend filtering. Although

they covered all orders of trend filtering in contrast to our method, their methods were however restricted to lattice designs. Moreover, they imposed weaker penalties on their models, which resulted in their estimators converging to the true underlying function at dimension-dependent rates. In Section 9 of the Supplementary Material (Ki, Fang and Guntuboyina (2024)), we describe the method of Ortelli and van de Geer (2021) and compare their estimator to the discrete formation of our estimator in the equally spaced lattice design setting (see Remark 9.4 of Ki, Fang and Guntuboyina (2024) for more details).

We have also mentioned that $V_{\mathrm{mars}}(\cdot)$ can be viewed as second-order Hardy–Krause variation (anchored at $\mathbf{0}$). Our estimation strategy can thus be seen as second-order Hardy–Krause variation denoising. In Fang, Guntuboyina and Sen (2021), first-order Hardy–Krause variation denoising (i.e., least squares estimation over functions with bounded Hardy–Krause variation) was studied. First-order Hardy–Krause variation denoising leads to piecewise constant fits while our method leads to MARS fits (linear combinations of products of ReLU functions of individual variables). Fang, Guntuboyina and Sen (2021) also proved that their estimator achieves a dimension-free (up to a logarithmic multiplicative factor) rate of convergence. However, it should be noted that their result is only proved in the fixed lattice design setting. Moreover, unlike our method, interaction order restriction is not considered in Fang, Guntuboyina and Sen (2021).

As is clear from the form of our functions (3) and our complexity measure (6), our method can also be considered as a multivariate ANOVA modeling method based on total variation constraints. There are a few works that utilize total variation penalties in multivariate ANOVA modeling. Petersen, Witten and Simon (2016), Yang and Tan (2018), and Sadhanala and Tibshirani (2019) utilized total variation of univatiate functions in additive modeling, which can be seen as a special case of ANOVA modeling where the interaction between covariates is not allowed. Also, Yang and Tan (2021) used for multivariate ANOVA modeling a class of penalties characterized in terms of certain hierarchical notions of total variation. Their hierarchical total variations are defined using a pre-fixed grid of points. Interestingly, for functions f that are sufficiently smooth, one of their hierarchical total variations (corresponding to $m = 2$ in their notation) converges to $V_{\mathrm{mars}}(f)$ as the grid resolution becomes arbitrarily small.

It should be mentioned that Lin (1998, 2000) also studied a multivariate ANOVA modeling method, but instead of L^1 norms as in our paper, they worked with penalties that are related to the squared L^2-Sobolev norms. Relevance of their works to our paper is therefore not from the type of penalties but from tensor product structures on their basis functions. As our basis functions (1) are the tensor products of univariate ReLU functions, their basis functions are also the tensor products of univariate functions whose smoothness is constrained by the L^2-Sobolev norms. It is notable that the multivariate function spaces considered in Lin (1998, 2000) are defined as an appropriate completion of the pre-Hilbert space given by the tensor product of the univariate L^2-Sobolev spaces. We are curious whether we can also view our function classes (e.g., $\mathcal{F}^{d,d}_{\infty-\mathrm{mars}}$) as an appropriate completion of a tensor product space. However, it is unclear to us at this point what norm should be chosen for completion as a counterpart of the L^2-Sobolev norms of Lin (1998, 2000). We believe this is an interesting direction to extend our work, which can provide a new perspective on our function spaces.

In addition, van der Laan, Benkeser and Cai (2023) discussed function classes similar to $\mathcal{F}^{d,d}_{\infty-\mathrm{mars}}$ and norms similar to $V_{\mathrm{mars}}(\cdot)$ and used them for estimation in settings that are different from our classical nonparametric regression framework.

6. 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 (13). For the usual MARS estimator, we use the R package earth based on Friedman (1991) and Friedman (1993).

6.1. *Simulation studies.* Mammen and van de Geer (1997) and Tibshirani (2014) 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 (second-order) 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 $X^{(i)}$ are uniformly distributed on $[0,1]^d$ and y_i are generated according to the model (20). For Function 1 and Function 3, we also consider the equally spaced lattice design where

$$\{x^{(1)},\dots,x^{(n)}\} = \prod_{k=1}^{d}\left\{\frac{i_k}{n_k} : i_k \in [0:(n_k-1)]\right\}$$

for some integers $n_k \geq 2$. 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 1.

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

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

  for $x_1, x_2 \in [0,1]$, where $r(x_1,x_2) = \sqrt{(x_1-0.3)^2+(x_2-0.4)^2}$. 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

$$f^*(x_1,x_2) = 5\sin\left(\frac{4}{\sqrt{x_1^2+x_2^2}+0.001}\right)+7.5$$

  for $x_1, x_2 \in [0,1]$. This function is a (scaled) two-dimensional version of the Doppler function used for simulation studies in Mammen and van de Geer (1997) and Tibshirani (2014). 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 Friedman (1991), Section 4.3 and 4.4), which have been frequently utilized to measure the performance of nonparametric regression methods (see, e.g., Meyer, Leisch and Hornik (2003), Potts and Schmischke (2021, 2022)). The four functions (Function 5, Function 6, Function 7, and Function 8) are defined as below, and for each function, we generate data $(X^{(1)}, y_1),\dots,(X^{(n)}, y_n)$ according to the model (20) where $X^{(i)}$ are uniformly distributed on $[0,1]^d$. For Function 5 and Function 6, we consider Gaussian errors with standard deviation 1 as in Friedman (1991),

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 Friedman (1991), Section 4.4 (see also Meyer, Leisch and Hornik (2003)). Also, for each function, we conduct experiments on three different sample sizes as in Friedman (1991): 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
$$f^*(x_1, x_2, x_3, x_4) = \sqrt{\big(T_1(x_1)\big)^2 + \big(T_2(x_2) \cdot x_3 - \big(T_2(x_2) \cdot T_4(x_4)\big)^{-1}\big)^2}$$
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
$$f^*(x_1, x_2, x_3, x_4) = \arctan\left(\frac{T_2(x_2) \cdot x_3 - (T_2(x_2) \cdot T_4(x_4))^{-1}}{T_1(x_1)}\right)$$
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 $\tilde{f}^{d,2}_{n,V}$ to the usual MARS estimator whose order of interaction is restricted to 2. The approximate method is considered only when explanatory variables are uniformly distributed. The positive integers N_k for the approximate method is 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 (17) for the fixed lattice design, and for the uniform design, we generate 1000 new samples and approximate the prediction error (22).

Table 1 presents the average loss of our estimators and the usual MARS estimator over 25 repetitions for each function. In Table 1, we can 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 1 also shows that our methods estimate the Friedman's functions F2, F3, and F4 much better, while the usual MARS method do 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 greatly improve the performance of our methods. We think it can be a promising avenue for future work.

Moreover, we can observe from Table 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

TABLE 1

The average and standard error of squared empirical L^2 norms (17) (lattice design) or prediction errors (22) (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	Number of data	Usual method	Original method	Approximate method
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)**

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 we observed 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.

6.2. *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 is presented in Section 10 of the Supplementary Material (Ki, Fang and Guntuboyina (2024)).

TABLE 2
The number of explanatory variables, the number of data, and the type of our method employed for each dataset. Original and Approx here stand for the original and approximate method, respectively

Dataset	Dimension	Number of data	Method
Earnings	2	25,437	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

For each dataset, we first linearly transform each explanatory variable into $[0, 1]$. In general, there can be multiple options for linear transformations. If the domain of an explanatory variable is known as $[m, M]$, then it is natural to subtract m from the variable 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 explanatory variables of order up to 2. In other words, we use $\hat{f}^{d,2}_{n,V}$ or its approximate version $\tilde{f}^{d,2}_{n,V}$ 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_k to 25. Furthermore, the tuning parameter V is chosen by 10-fold cross-validation in all cases.

In Table 2, we present for each dataset the number of explanatory variables, the number of data, and the type of our method we use (whether the original or the approximate method). Table 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, 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

TABLE 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)**

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.

Acknowledgments. We are immensely grateful to the anonymous referees for their constructive comments and suggestions, which significantly improved the quality of the paper. We also thank Prof. Trevor Hastie for helpful comments and discussion.

Funding. The first author was supported by NSF Grant DMS-2023505, NSF CAREER Grant DMS-1654589, and NSF Grant DMS-2210504.

The third author was supported by NSF CAREER Grant DMS-1654589 and NSF Grant DMS-2210504.

SUPPLEMENTARY MATERIAL

Supplement to "MARS via LASSO" (DOI: 10.1214/24-AOS2384SUPPA; .pdf). It contains a brief introduction to Hardy–Krause variation; the results of Section 4 for the case $d = s = 2$; comparison between our method in the equally spaced lattice design setting and the multivariate trend filtering method of Ortelli and van de Geer (2021); brief descriptions of the datasets used for the numerical experiments in Section 6; and proofs of all our results.

Code (DOI: 10.1214/24-AOS2384SUPPB; .zip). It contains the code for all the numerical experiments in Section 6. The code is also available at https://github.com/DohyeongKi/mars-lasso-paper.

REFERENCES

AISTLEITNER, C. and DICK, J. (2015). Functions of bounded variation, signed measures, and a general Koksma–Hlawka inequality. *Acta Arith.* **167** 143–171. MR3312093 https://doi.org/10.4064/aa167-2-4

ARTSTEIN, S., MILMAN, V., SZAREK, S. and TOMCZAK-JAEGERMANN, N. (2004). On convexified packing and entropy duality. *Geom. Funct. Anal.* **14** 1134–1141. MR2105957 https://doi.org/10.1007/s00039-004-0486-3

BLEI, R., GAO, F. and LI, W. V. (2007). Metric entropy of high dimensional distributions. *Proc. Amer. Math. Soc.* **135** 4009–4018. MR2341952 https://doi.org/10.1090/S0002-9939-07-08935-6

BREDIES, K. and PIKKARAINEN, H. K. (2013). Inverse problems in spaces of measures. *ESAIM Control Optim. Calc. Var.* **19** 190–218. MR3023066 https://doi.org/10.1051/cocv/2011205

BUNGARTZ, H.-J. and GRIEBEL, M. (2004). Sparse grids. *Acta Numer.* **13** 147–269. MR2249147 https://doi.org/10.1017/S0962492904000182

CANDÈS, E. J. and FERNANDEZ-GRANDA, C. (2014). Towards a mathematical theory of super-resolution. *Comm. Pure Appl. Math.* **67** 906–956. MR3193963 https://doi.org/10.1002/cpa.21455

CHEN, X. and LI, W. V. (2003). Quadratic functionals and small ball probabilities for the m-fold integrated Brownian motion. *Ann. Probab.* **31** 1052–1077. MR1964958 https://doi.org/10.1214/aop/1048516545

CONDAT, L. (2020). Atomic norm minimization for decomposition into complex exponentials and optimal transport in Fourier domain. *J. Approx. Theory* **258** 105456, 24. MR4127283 https://doi.org/10.1016/j.jat.2020.105456

DE CASTRO, Y. and GAMBOA, F. (2012). Exact reconstruction using Beurling minimal extrapolation. *J. Math. Anal. Appl.* **395** 336–354. MR2943626 https://doi.org/10.1016/j.jmaa.2012.05.011

DE CASTRO, Y., GAMBOA, F., HENRION, D. and LASSERRE, J.-B. (2017). Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *IEEE Trans. Inf. Theory* **63** 621–630. MR3599963 https://doi.org/10.1109/TIT.2016.2619368

DENOYELLE, Q., DUVAL, V., PEYRÉ, G. and SOUBIES, E. (2020). The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. *Inverse Probl.* **36** 014001, 42. MR4040984 https://doi.org/10.1088/1361-6420/ab2a29

DŨNG, D., TEMLYAKOV, V. and ULLRICH, T. (2018). *Hyperbolic Cross Approximation*. *Advanced Courses in Mathematics*. *CRM Barcelona*. Birkhäuser, Cham. Edited and with a foreword by Sergey Tikhonov. MR3887571 https://doi.org/10.1007/978-3-319-92240-9

DUNKER, T., LINDE, W., KÜHN, T. and LIFSHITS, M. A. (1999). Metric entropy of integration operators and small ball probabilities for the Brownian sheet. *J. Approx. Theory* **101** 63–77. MR1724026 https://doi.org/10.1006/jath.1999.3354

DUVAL, V. and PEYRÉ, G. (2015). Exact support recovery for sparse spikes deconvolution. *Found. Comput. Math.* **15** 1315–1355. MR3394712 https://doi.org/10.1007/s10208-014-9228-6

FANG, B., GUNTUBOYINA, A. and SEN, B. (2021). Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *Ann. Statist.* **49** 769–792. MR4255107 https://doi.org/10.1214/20-aos1977

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–67. MR1091842 https://doi.org/10.1214/aos/1176347963

FRIEDMAN, J. H. (1993). Fast MARS Technical Report No. LCS110, Stanford Univ. Press, Stanford.

GAO, F. (2008). Entropy estimate for k-monotone functions via small ball probability of integrated Brownian motion. *Electron. Commun. Probab.* **13** 121–130. MR2386068 https://doi.org/10.1214/ECP.v13-1357

GUNTUBOYINA, A., LIEU, D., CHATTERJEE, S. and SEN, B. (2020). Adaptive risk bounds in univariate total variation denoising and trend filtering. *Ann. Statist.* **48** 205–229. MR4065159 https://doi.org/10.1214/18-AOS1799

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

KI, D., FANG, B. and GUNTUBOYINA, A. (2024). Supplement to "MARS via LASSO." https://doi.org/10.1214/24-AOS2384SUPPA, https://doi.org/10.1214/24-AOS2384SUPPB

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 https://doi.org/10.1137/070690274

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and processes*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Springer, Berlin. MR1102015 https://doi.org/10.1007/978-3-642-20212-4

LI, W. V. and LINDE, W. (1999). Approximation, metric entropy and small ball estimates for Gaussian measures. *Ann. Probab.* **27** 1556–1578. MR1733160 https://doi.org/10.1214/aop/1022677459

LIN, Y. (1998). *Tensor Product Space ANOVA Models in Multivariate Function Estimation*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of Pennsylvania. MR2697355

LIN, Y. (2000). Tensor product space ANOVA models. *Ann. Statist.* **28** 734–755. MR1792785 https://doi.org/10.1214/aos/1015951996

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931 https://doi.org/10.1214/aos/1034276635

MEYER, D., LEISCH, F. and HORNIK, K. (2003). The support vector machine under test. *Neurocomputing* **55** 169–186.

NEMIROVSKI, A. S. (2000). Topics in nonparametric statistics. In *Lectures on Probability Theory and Statistics*: *École D'Été de Probabilités de Saint-Flour XXVIII*-1998. *Lecture Notes in Mathematics* **1738**. Springer, Berlin, Heidelberg.

ORTELLI, F. and VAN DE GEER, S. (2021). Tensor denoising with trend filtering. *Math. Stat. Learn.* **4** 87–142. MR4383732 https://doi.org/10.4171/msl/26

OWEN, A. B. (2005). Multidimensional variation for quasi-Monte Carlo. In *Contemporary Multivariate Analysis and Design of Experiments*. *Ser. Biostat.* **2** 49–74. World Scientific, Hackensack, NJ. MR2271076

PARHI, R. and NOWAK, R. D. (2021). Banach space representer theorems for neural networks and ridge splines. *J. Mach. Learn. Res.* **22** 1–40. MR4253736

PARHI, R. and NOWAK, R. D. (2023). Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Trans. Inf. Theory* **69** 1125–1140. MR4564646 https://doi.org/10.1109/tit.2022.3208653

PETERSEN, A., WITTEN, D. and SIMON, N. (2016). Fused lasso additive model. *J. Comput. Graph. Statist.* **25** 1005–1025. MR3572026 https://doi.org/10.1080/10618600.2015.1073155

POTTS, D. and SCHMISCHKE, M. (2021). Interpretable approximation of high-dimensional data. *SIAM J. Math. Data Sci.* **3** 1301–1323. MR4344888 https://doi.org/10.1137/21M1407707

POTTS, D. and SCHMISCHKE, M. (2022). Learning multivariate functions with low-dimensional structures using polynomial bases. *J. Comput. Appl. Math.* **403** Paper No. 113821, 19. MR4321625 https://doi.org/10.1016/j.cam.2021.113821

ROSSET, S., SWIRSZCZ, G., SREBRO, N. and ZHU, J. (2007). ℓ_1 regularization in infinite dimensional feature spaces. In *Learning Theory*. *Lecture Notes in Computer Science* **4539** 544–558. Springer, Berlin. MR2397611 https://doi.org/10.1007/978-3-540-72927-3_39

RUDIN, W. (1987). *Real and Complex Analysis*, 3rd ed. McGraw-Hill, New York. MR0924157

SADHANALA, V. and TIBSHIRANI, R. J. (2019). Additive models with trend filtering. *Ann. Statist.* **47** 3032–3068. MR4025734 https://doi.org/10.1214/19-AOS1833

SADHANALA, V., WANG, Y.-X., HU, A. J. and TIBSHIRANI, R. J. (2021). Multivariate trend filtering for lattice data. arXiv preprint. Available at arXiv:2112.14758.

STEIDL, G., DIDAS, S. and NEUMANN, J. (2006). Splines in higher order TV regularization. *Int. J. Comput. Vis.* **70** 241–255.

TEMLYAKOV, V. (2018). *Multivariate Approximation*. *Cambridge Monographs on Applied and Computational Mathematics* **32**. Cambridge Univ. Press, Cambridge. MR3837133 https://doi.org/10.1017/9781108689687

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487 https://doi.org/10.1214/13-AOS1189

VAN DER LAAN, M. J., BENKESER, D. and CAI, W. (2023). Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. *Int. J. Biostat.* **19** 261–289.

YANG, T. and TAN, Z. (2018). Backfitting algorithms for total-variation and empirical-norm penalized additive modelling with high-dimensional data. *Stat* **7** e198, 19. MR3905854 https://doi.org/10.1002/sta4.198

YANG, T. and TAN, Z. (2021). Hierarchical total variations and doubly penalized ANOVA modeling for multivariate nonparametric regression. *J. Comput. Graph. Statist.* **30** 848–862. MR4356590 https://doi.org/10.1080/10618600.2021.1923513