The Annals of Statistics
2021, Vol. 49, No. 2, 769–792
https://doi.org/10.1214/20-AOS1977

MULTIVARIATE EXTENSIONS OF ISOTONIC REGRESSION AND TOTAL VARIATION DENOISING VIA ENTIRE MONOTONICITY AND HARDY–KRAUSE VARIATION

BY BILLY FANG[1,*], ADITYANAND GUNTUBOYINA[1,†] AND BODHISATTVA SEN[2]

[1]*Department of Statistics, University of California, [*]blfang@berkeley.edu; [†]aditya@stat.berkeley.edu*
[2]*Department of Statistics, Columbia University, bodhi@stat.columbia.edu*

We consider the problem of nonparametric regression when the covariate is d dimensional, where $d \geq 1$. In this paper, we introduce and study two nonparametric least squares estimators (LSEs) in this setting—the entirely monotonic LSE and the constrained Hardy–Krause variation LSE. We show that these two LSEs are natural generalizations of univariate isotonic regression and univariate total variation denoising, respectively, to multiple dimensions. We discuss the characterization and computation of these two LSEs obtained from n data points. We provide a detailed study of their risk properties under the squared error loss and fixed uniform lattice design. We show that the finite sample risk of these LSEs is always bounded from above by $n^{-2/3}$ modulo logarithmic factors depending on d; thus these nonparametric LSEs avoid the curse of dimensionality to some extent. We also prove nearly matching minimax lower bounds. Further, we illustrate that these LSEs are particularly useful in fitting rectangular piecewise constant functions. Specifically, we show that the risk of the entirely monotonic LSE is almost parametric (at most $1/n$ up to logarithmic factors) when the true function is well approximable by a rectangular piecewise constant entirely monotone function with not too many constant pieces. A similar result is also shown to hold for the constrained Hardy–Krause variation LSE for a simple subclass of rectangular piecewise constant functions. We believe that the proposed LSEs yield a novel approach to estimating multivariate functions using convex optimization that avoid the curse of dimensionality to some extent.

1. Introduction.

Consider the problem of nonparametric regression where the goal is to estimate an unknown regression function $f^* : [0, 1]^d \to \mathbb{R}$ $(d \geq 1)$ from noisy observations at fixed design points $\mathbf{x}_1, \dots, \mathbf{x}_n \in [0, 1]^d$. Specifically, we observe responses $y_1, \dots, y_n$ drawn according to the model

$$y_i = f^*(\mathbf{x}_i) + \xi_i, \quad \text{where } \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \text{ for } i = 1, \dots, n, \tag{1}$$

$\sigma^2 > 0$ is unknown, and the purpose is to nonparametrically estimate f^* known to belong to a prespecified function class. In the univariate ($d = 1$) case, two such important function classes are: (i) the class of *monotone nondecreasing* functions in which case f^* is usually estimated by the isotonic least squares estimator (LSE) (see, e.g., Robertson et al. [46], Groeneboom and Jongbloed [26], Barlow et al. [3], Brunk [8], Ayer et al. [2]) and (ii) the class of functions whose *total variation* is bounded by a specific constant in which case it is natural to estimate f^* by total variation denoising (see, e.g., Rudin et al. [47], Mammen and van de Geer [37], Chambolle et al. [10], Condat [14]). Both these estimators—isotonic regression and total variation denoising—have a long history and are very well studied. For example, it is known

Received March 2019; revised December 2019.
MSC2020 subject classifications. 62G08.
Key words and phrases. Almost parametric risk, bounded mixed derivative, (constrained) least squares estimation, curse of dimensionality, dimension independent risk, multivariate shape constrained regression, nonparametric regression, risk under the squared error loss.

that both these estimators produce piecewise constant fits and have finite sample risk (under the squared error loss) bounded from above by a constant multiple of $n^{-2/3}$ (see, e.g., Meyer and Woodroofe [38], Zhang [63], Mammen and van de Geer [37]). Moreover, it is well known that both these estimators are especially useful in fitting piecewise constant functions where their risk is almost parametric (at most $1/n$ up to logarithmic factors); see, for example, Guntuboyina and Sen [27], Dalalyan et al. [16] and Guntuboyina et al. [28] and the references therein.

In this paper, we try to answer the following question: "What is a natural generalization of univariate isotonic regression and univariate total variation denoising to multiple dimensions?" To answer this question, we introduce and study two (constrained) LSEs for estimating $f^* : [0,1]^d \to \mathbb{R}$ where $d \geq 1$. We show that both these LSEs yield rectangular piecewise constant fits and have finite sample risk that is bounded from above by $n^{-2/3}$ (modulo logarithmic factors depending on d), thereby avoiding the curse of dimensionality to some extent. Further, we study the characterization and computation of these two estimators: the LSEs are obtained as solutions to convex optimization problems—in fact, quadratic programs with linear constraints—and are thus easily computable. Moreover, as in the case $d = 1$, we illustrate that these LSEs are particularly useful in fitting rectangular piecewise constant functions and can have almost parametric risk (up to logarithmic factors). These results are directly analogous to the univariate results mentioned in the previous paragraph, and thus justify our claim that our proposed estimators are natural multivariate generalizations of univariate isotonic regression and univariate total variation denoising.

Our first estimator is the LSE over $\mathcal{F}^d_{\mathrm{EM}}$, the class of *entirely monotone* functions on $[0,1]^d$:

$$\hat{f}_{\mathrm{EM}} \in \operatorname*{argmin}_{f \in \mathcal{F}^d_{\mathrm{EM}}} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(\mathbf{x}_i)\big)^2. \tag{2}$$

The class $\mathcal{F}^d_{\mathrm{EM}}$ of entirely monotone functions is formally defined in Section 2. Entire monotonicity is an existing generalization in multivariate analysis of the univariate notion of monotonicity (see, e.g., [1, 31, 34, 62]). Indeed, in the univariate case when $d = 1$, the class $\mathcal{F}^1_{\mathrm{EM}}$ is precisely the class of nondecreasing functions on $[0,1]$, and thus, for $d = 1$, the estimator (2) reduces to the usual isotonic LSE. For $d = 2$, the class $\mathcal{F}^2_{\mathrm{EM}}$ consists of all functions $f : [0,1]^2 \to \mathbb{R}$ which satisfy both $f(a_1, a_2) \leq f(b_1, b_2)$ and

$$f(b_1, b_2) - f(a_1, b_2) - f(b_1, a_2) + f(a_1, a_2) \geq 0, \tag{3}$$

for every $0 \leq a_1 \leq b_1 \leq 1$ and $0 \leq a_2 \leq b_2 \leq 1$. The formal definition of $\mathcal{F}^d_{\mathrm{EM}}$ for general $d \geq 1$ is given in Section 2. We remark that in general, entire monotonicity is different from the usual notion of monotonicity in classical multivariate isotonic regression [46]; see Lemma 2.1 for a connection between these two notions. We also remark that $\mathcal{F}^d_{\mathrm{EM}}$ is closed under translation and nonnegative scaling; that is, if $f \in \mathcal{F}^d_{\mathrm{EM}}$, then $af + b \in \mathcal{F}^d_{\mathrm{EM}}$ for any $a \geq 0$ and $b \in \mathbb{R}$. Additionally, the collection of right-continuous functions in $\mathcal{F}^d_{\mathrm{EM}}$ is precisely the collection of cumulative distribution functions of nonnegative measures on $[0,1]^d$ (see Lemma 2.2).

Our terminology of entire monotonicity is taken from Young and Young [62]. As a word of caution, we note that some authors (e.g., Aistleitner and Dick [1]) use the term "completely monotone" in place of "entirely monotone." We use the latter terminology because "completely monotone" has been used in the literature for other notions (see, e.g., [22, 24, 60]) which are unrelated to our definition of entire monotonicity. Entire monotonicity has also been referred by other names in the literature (e.g., it has been referred to as "quasi-monotone" in Hobson [31]).

The second main estimator that we study in this paper involves $V_{\mathrm{HK}\mathbf{0}}(\cdot)$, the *variation in the sense of Hardy and Krause (anchored at* $\mathbf{0}$*)*, which we shorten to *Hardy–Krause variation* or HK**0** *variation*. The HK**0** variation of a univariate function $f : [0,1] \to \mathbb{R}$ is simply the total variation of the function, that is,

$$V_{\mathrm{HK}\mathbf{0}}(f) = \sup_{0=x_0<x_1<\cdots<x_k=1} \sum_{i=0}^{k-1} |f(x_{i+1}) - f(x_i)|, \tag{4}$$

where the supremum is over all $k \geq 1$ and all partitions $0 = x_0 < x_1 < \cdots < x_k = 1$ of $[0,1]$. Thus HK**0** variation is a generalization of one-dimensional total variation to multiple dimensions. For $d = 2$, HK**0** variation is defined in the following way: for $f : [0,1]^2 \to \mathbb{R}$,

$$\begin{aligned} V_{\mathrm{HK}\mathbf{0}}(f) &:= V_{\mathrm{HK}\mathbf{0}}\big(x \mapsto f(x,0)\big) + V_{\mathrm{HK}\mathbf{0}}\big(x \mapsto f(0,x)\big) \\ &\quad + \sup \sum_{0 \leq l_1 < k_1, 0 \leq l_2 < k_2} \big| f\big(x^{(1)}_{l_1+1}, x^{(2)}_{l_2+1}\big) - f\big(x^{(1)}_{l_1}, x^{(2)}_{l_2+1}\big) \\ &\quad - f\big(x^{(1)}_{l_1+1}, x^{(2)}_{l_2}\big) + f\big(x^{(1)}_{l_1}, x^{(2)}_{l_2}\big)\big|, \end{aligned} \tag{5}$$

where the first two terms in the right-hand side above are defined via the univariate definition (4) and the supremum in the third term above is over all pairs of partitions $0 = x^{(1)}_0 < x^{(1)}_1 < \cdots < x^{(1)}_{k_1} = 1$ and $0 = x^{(2)}_0 < x^{(2)}_1 < \cdots < x^{(2)}_{k_2} = 1$ of $[0,1]$. Note that a special role is played in the first two terms of the right-hand side of (5) by the point $(0,0)$ and this is the reason for the phrase "anchored at **0**." For smooth functions $f : [0,1]^2 \to \mathbb{R}$, it can be shown that

$$V_{\mathrm{HK}\mathbf{0}}(f) = \int_0^1 \int_0^1 \left| \frac{\partial^2 f}{\partial x_1 \partial x_2} \right| dx_1\, dx_2 + \int_0^1 \left| \frac{\partial f(\cdot,0)}{\partial x_1} \right| dx_1 + \int_0^1 \left| \frac{\partial f(0,\cdot)}{\partial x_2} \right| dx_2$$

and, from the first term in the right-hand side above, it is clear that the HK**0** variation is related to the L^1 norm of the mixed derivative. The definition of HK**0** variation for general $d \geq 1$ is given in Section 2. HK**0** variation is quite different from the usual definition of multivariate total variation (see, e.g., Ziemer [65],Chapter 5) as explained briefly in Section 2.

Functions that are piecewise constant on axis-aligned rectangular pieces (see Definition 2.3) have finite HK**0** variation as explained in Section 2. More generally, the collection of right-continuous functions of finite HK**0** variation is precisely the same as the collection of cumulative distribution functions of finite signed measures (see Lemma 2.5). An example of a function with infinite HK**0** variation is the indicator function of an open d-dimensional ball contained in $[0,1]^d$ (see [44], Section 12).

Our second estimator is the constrained LSE over functions with HK**0** variation bounded by some tuning parameter $V > 0$:

$$\hat{f}_{\mathrm{HK}\mathbf{0},V} \in \underset{f : V_{\mathrm{HK}\mathbf{0}}(f) \leq V}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \big(y_i - f(\mathbf{x}_i)\big)^2. \tag{6}$$

This estimator is a generalization of total variation denoising to $d \geq 2$ because in the case $d = 1$, HK**0** variation coincides with total variation, and thus, the above estimator performs univariate total variation denoising, sometimes also called trend filtering of first order [10, 14, 33, 37, 47, 53]. This generalization is different from the usual multivariate total variation denoising as in Rudin et al. [47] (see Section 5 for more discussion on how $\hat{f}_{\mathrm{HK}\mathbf{0},V}$ is different from the multivariate total variation regularized estimator). It is also possible to define the HK**0** variation estimator in the following penalized form:

$$\hat{f}_{\mathrm{HK}\mathbf{0},\lambda} \in \underset{f}{\operatorname{argmin}} \frac{1}{n} \left\{ \sum_{i=1}^n \big(y_i - f(\mathbf{x}_i)\big)^2 + \lambda V_{\mathrm{HK}\mathbf{0}}(f) \right\} \tag{7}$$

for a tuning parameter $\lambda > 0$. In this paper, we shall focus on the constrained form in (6) although analogues of our results for the penalized estimator (7) can also be proved.

Before proceeding further, let us note that entire monotonicity is related to HK$\mathbf{0}$ variation in much the same way as univariate monotonicity is related to univariate total variation. Indeed, for functions in one variable, the following two properties are well known:

1. Every function $f : [0, 1] \to \mathbb{R}$ of bounded variation can be written as the difference of two monotone functions $f = f_+ - f_-$ and the total variation of f equals the sum of the variations of f_+ and f_-.
2. If $f : [0, 1] \to \mathbb{R}$ is nondecreasing, then its total variation on $[0, 1]$ is simply $f(1) - f(0)$.

These two facts generalize almost verbatim to entire monotonicity and HK$\mathbf{0}$ variation (see Lemma 2.4). Thus, in some sense, entire monotonicity is to Hardy–Krause variation as monotonicity is to total variation.

Although the terminology of "entire monotonicity" does not seem to have been used previously in the statistics literature, entirely monotone functions are closely related to cumulative distribution functions of nonnegative measures which appear routinely in statistics. HK$\mathbf{0}$ variation has appeared previously in statistics in the literature on quasi-Monte Carlo (see, e.g., [29, 44]) as well as in the power analysis of certain sequential detection problems (see, e.g., [45]). Additionally Benkeser and Van Der Laan [6] (see also [54–57]) considered the class $\{f : V_{\mathrm{HK}\mathbf{0}}(f) \le V\}$ in their "highly adaptive LASSO" estimator and exploited its connections to the LASSO in a setting that is different from our classical nonparametric regression framework. They also used the terminology of "sectional variation norm" to refer to the Hardy–Krause variation (see also [25], Section 2). An estimator very similar to (6) was proposed by Mammen and van de Geer [37] for $d = 2$ when the design points take values in a uniformly spaced grid (this estimator of [37] is described in Section 3.1). Also, Lin [35] proposed an estimator in the context of the Gaussian white noise model that bears some similarities to (6) (this connection is detailed in Section 5).

The goal of this paper is to analyze the properties of the estimators (2) and (6). Here is a description of our main results. Section 3 concerns the computation of these estimators. Note that, as stated, the optimization problems defining our estimators (2) and (6) are convex (albeit infinite-dimensional). We show that, given arbitrary data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$, the two estimators (2) and (6) can be computed by solving a nonnegative least squares (NNLS) problem and a LASSO problem, respectively, with a suitable design matrix that only depends on the design-points $\mathbf{x}_1, \dots, \mathbf{x}_n$. It is interesting to note that the design matrices in the two finite-dimensional problems for computing (2) and (6) are exactly the same. Our main results in this section (Proposition 3.1 and Proposition 3.3) imply that $\widehat{f}_{\mathrm{EM}}$ and $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ can be taken to be of the form

$$\widehat{f}_{\mathrm{EM}} = \sum_{j=1}^{p} (\widehat{\beta}_{\mathrm{EM}})_j \cdot \mathbb{I}_{[\mathbf{z}_j,\mathbf{1}]} \quad \text{and} \quad \widehat{f}_{\mathrm{HK}\mathbf{0},V} = \sum_{j=1}^{p} (\widehat{\beta}_{\mathrm{HK}\mathbf{0},V})_j \cdot \mathbb{I}_{[\mathbf{z}_j,\mathbf{1}]} \tag{8}$$

for some $\mathbf{z}_1, \dots, \mathbf{z}_p$ that only depend on the design points $\mathbf{x}_1, \dots, \mathbf{x}_n$ and vectors $\widehat{\boldsymbol{\beta}}_{\mathrm{EM}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{HK}\mathbf{0},V}$ in $\mathbb{R}^p$ which are obtained by solving the NNLS problem (24) and the LASSO problem (26), respectively. Here, $\mathbb{I}_{[\mathbf{z}_j,\mathbf{1}]}$ denotes the indicator of the rectangle $[\mathbf{z}_j, \mathbf{1}]$ (defined via (16)). Because NNLS and LASSO typically lead to sparse solutions, the vectors $\widehat{\boldsymbol{\beta}}_{\mathrm{EM}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{HK}\mathbf{0},V}$ will be sparse which clearly implies that $\widehat{f}_{\mathrm{EM}}$ and $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ as given above (8) will be piecewise constant on axis-aligned rectangles. Therefore, our estimators give rectangular piecewise constant fits to data and this generalizes the fact that univariate isotonic regression and total variation denoising yield piecewise constant fits. In the case when the design points $\mathbf{x}_1, \dots, \mathbf{x}_n$ form an equally spaced lattice in $[0, 1]^d$ (see the definition (30) for the precise formulation of this assumption), the points $\mathbf{z}_1, \dots, \mathbf{z}_p$ can simply be taken to be $\mathbf{x}_1, \dots, \mathbf{x}_n$

and, in this case, more explicit expressions can be given for the estimators (see Section 3.1 for details). It should be noted that the lattice design is quite commonly used for theoretical studies in multidimensional nonparametric function estimation (see, e.g., [39]) especially in connection with image analysis (see, e.g., [10, 15]).

We also investigate the accuracy properties of $\widehat{f}_{\mathrm{EM}}$ and $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ via the study of their risk behavior under the standard fixed design squared error loss function. Specifically, we define the risk of an estimator $\widehat{f}$ by

$$(9)\qquad \mathcal{R}(\widehat{f}, f^*) := \mathbb{E}\mathcal{L}(\widehat{f}, f^*) \quad \text{where } \mathcal{L}(\widehat{f}, f^*) := \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2.$$

We prove results on the risk of $\widehat{f}_{\mathrm{EM}}$ and $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ in the case of the aforementioned lattice design. In this setting, our main results are described below.

We analyze the risk of $\widehat{f}_{\mathrm{EM}}$ under the (well-specified) assumption that $f^* \in \mathcal{F}^d_{\mathrm{EM}}$. We prove in Theorem 4.1 that, for $n \geq 1$,

$$(10)\qquad \mathcal{R}(\widehat{f}_{\mathrm{EM}}, f^*) \leq \frac{C(d, \sigma, V^*)}{n^{2/3}}(\log(en))^{\frac{2d-1}{3}},$$

where

$$V^* := f^*(1, \dots, 1) - f^*(0, \dots, 0)$$

and $C(d, \sigma, V^*)$ depends only on d, σ and V^* (see statement of Theorem 4.1 for the explicit form of $C(d, \sigma, V^*)$). Note that the dimension d appears in (10) only through the logarithmic term which means that we obtain "dimension independent rates" ignoring logarithmic factors. Some intuition for why the constraint of entire monotononicity is able to mitigate the usual curse of dimensionality is provided in Section 5. Other nonparametric estimators exhibiting such dimension independent rates can be found in [4, 13, 35, 40, 50, 59]. In Theorem 4.2, we prove a minimax lower bound which implies that the dependence on d through the logarithmic term in (10) cannot be avoided for any estimator.

We also prove in Theorem 4.4 that $\mathcal{R}(\widehat{f}_{\mathrm{EM}}, f^*)$ is smaller than the bound given by (10) when $f^* \in \mathcal{F}^d_{\mathrm{EM}}$ is *rectangular piecewise constant*. Loosely speaking, we say that $f : [0, 1]^d \to \mathbb{R}$ is rectangular piecewise constant if it is constant on each set in a partition of $[0, 1]^d$ into axis-aligned rectangles and the smallest cardinality of such a partition shall be denoted by $k(f)$ (see Definition 2.3 for the precise definitions). In Theorem 4.4, we prove that whenever $f^* \in \mathcal{F}^d_{\mathrm{EM}}$ is rectangular piecewise constant, we have

$$(11)\qquad \mathcal{R}(\widehat{f}_{\mathrm{EM}}, f^*) \leq C_d\sigma^2\frac{k(f^*)}{n}(\log(en))^{\frac{3d}{2}}(\log(e\log(en)))^{\frac{2d-1}{2}}$$

for a positive constant C_d which only depends on d. Note that when $k(f^*)$ is not too large, the right-hand side of (11) converges to zero as $n \to \infty$ at a faster rate compared to the right-hand side of (10). Thus rectangular piecewise constant functions which also satisfy the constraint of entire monotonicity are estimated at nearly the parametric rate (ignoring the logarithmic factor) by the LSE $\widehat{f}_{\mathrm{EM}}$.

Let us now describe our results for the other estimator $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$. In Theorem 4.5, we prove that when $V_{\mathrm{HK}\mathbf{0}}(f^*) \leq V$ (note that V is the tuning parameter in the definition of $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$), then

$$(12)\qquad \mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, f^*) \leq \frac{C(d, \sigma, V)}{n^{2/3}}(\log(en))^{\frac{2d-1}{3}}.$$

Note that the right-hand sides of the bounds (12) and (10) are the same, and thus the estimator $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ also achieves dimension independent rates (ignoring logarithmic factors) (see

Section 5 for an explanation of this phenomenon). We also prove a minimax lower bound in Theorem 4.6 which implies that the dependence on d in the logarithmic term in (12) cannot be completely removed for any estimator.

In univariate total variation denoising, it is known that one obtains faster rates than given by the bound (12) when $f^* : [0, 1] \to \mathbb{R}$ is piecewise constant with not too many pieces. Indeed if f^* is piecewise constant for $d = 1$ with $k(f^*)$ pieces, then it has been proved that

$$\mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, f^*) \le C(c)\sigma^2 \frac{k(f^*)}{n} \log(en) \tag{13}$$

provided $V = V_{\mathrm{HK}\mathbf{0}}(f^*)$ and f^* satisfies a minimum length condition in that each constant piece has length at least $c/k(f^*)$ (the multiplicative term $C(c)$ in (13) only depends on this c appearing in the minimum length condition). A proof of this result can be found in [28], Corollary 2.3, and, for other similar results, see [16, 36, 41, 64]. In light of this univariate result, it is plausible to expect a bound similar to (11) for $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ when f^* is an axis-aligned rectangular piecewise constant function provided that the tuning parameter V is taken to be equal to $V_{\mathrm{HK}\mathbf{0}}(f^*)$ and provided that f^* satisfies a minimum length condition. We prove such a result for a class of simple rectangular piecewise constant functions $f^* : [0, 1]^d \to \mathbb{R}$ of the form

$$f^*(\cdot) = a_1 \mathbb{I}_{[\mathbf{x}^*, \mathbf{1}]}(\cdot) + a_0 \tag{14}$$

for some $a_1, a_0 \in \mathbb{R}$ and $\mathbf{x}^* \in [0, 1]^d$ (here $\mathbb{I}$ stands for the indicator function). It is easy to see that (14) represents a rectangular piecewise constant function with $k(f^*) \le 2^d$. In Theorem 4.7, we prove that when f^* is of the above form (14), then

$$\mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, f^*) \le C(c, d) \frac{\sigma^2}{n} (\log(en))^{\frac{3d}{2}} (\log(e \log(en)))^{\frac{2d-1}{2}} \tag{15}$$

provided the tuning parameter V equals $V_{\mathrm{HK}\mathbf{0}}(f^*)$ and $\mathbf{x}^* \in [0, 1]^d$ satisfies a *minimum size condition* (43). This latter condition, which is analogous to the minimum length condition in the univariate case, involves a positive constant c and the constant $C(c, d)$ appearing in (15) only depends on c and the dimension d. In the specific case when $d = 2$, the minimum length condition (43) can be weakened, as discussed in Appendix A of the Supplementary Material [21].

We are unable to prove versions of (15) for more general rectangular piecewise constant functions. However, some results in that direction have been proved in a very recent paper by Ortelli and van de Geer [41]. Their results are of a different flavor as they work with a similar but different estimator and a smaller loss function. Their proof techniques are also completely different from ours.

The rest of the paper is organized as follows. The notions of entire monotonicity and Hardy–Krause variation are formally defined for arbitrary $d \ge 1$ in Section 2 where we also collect some of their relevant properties. In Section 3, we discuss the computational aspects for solving the optimization problems in (2) and (6). The risk results for $\widehat{f}_{\mathrm{EM}}$ are given in Section 4.1 while the risk bounds for $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ are in Section 4.2. We discuss the connections of our contributions with other related work in Section 5. Due to space constraints, all the proofs are moved to the Supplementary Material [21]. The proofs for our risk results are given in Appendix C while the proofs of the results in Section 2 and Section 3 are given in Appendix D of [21]. Additional technical results used in the proofs of Appendix C are proved in Appendix E of [21]. The supplement also contains another risk bound for $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ (Appendix A), as well a section for simulations (Appendix B) that contains some examples and depictions of the two estimators, including an application to estimation in the bivariate current status model.

2. Entire monotonicity and Hardy–Krause variation. The aim of this section is to provide formal definitions of entire monotonicity and HK**0** variation for the convenience of the reader. We roughly follow the notation of Aistleitner and Dick [1] and Owen [44].

Let us first introduce some basic notation that will be used throughout the paper. We let $\mathbf{0} = (0, \dots, 0)$ and $\mathbf{1} = (1, \dots, 1)$. Given an integer m, we take $[m] := \{1, \dots, m\}$. For two points $\mathbf{a} = (a_1, \dots, a_d)$ and $\mathbf{b} = (b_1, \dots, b_d) \in [0, 1]^d$, we write

$$\mathbf{a} \prec \mathbf{b} \quad \text{if and only if} \quad a_j < b_j \quad \text{for every } j = 1, \dots, d$$

and

$$\mathbf{a} \preceq \mathbf{b} \quad \text{if and only if} \quad a_j \le b_j \quad \text{for every } j = 1, \dots, d.$$

When $\mathbf{a} \preceq \mathbf{b}$, we write

$$\begin{aligned} &[\mathbf{a}, \mathbf{b}] := \{\mathbf{x} : \mathbf{a} \preceq \mathbf{x} \preceq \mathbf{b}\} := \prod_{j=1}^{d} [a_j, b_j], \\ &[\mathbf{a}, \mathbf{b}) := \{\mathbf{x} : \mathbf{a} \preceq \mathbf{x} \prec \mathbf{b}\} := \prod_{j=1}^{d} [a_j, b_j). \end{aligned} \tag{16}$$

Note that $[\mathbf{a}, \mathbf{b}]$ is a closed axis-aligned rectangle and it has nonempty interior when $\mathbf{a} \prec \mathbf{b}$.

Given a function $f : [0, 1]^d \to \mathbb{R}$ and two distinct points $\mathbf{a} = (a_1, \dots, a_d)$, $\mathbf{b} = (b_1, \dots, b_d) \in [0, 1]^d$ with $\mathbf{a} \preceq \mathbf{b}$, we define the *quasi-volume* $\Delta(f; [\mathbf{a}, \mathbf{b}])$ by

$$\sum_{j_1=0}^{J_1} \cdots \sum_{j_d=0}^{J_d} (-1)^{j_1+\cdots+j_d} f\big(b_1 + j_1(a_1 - b_1), \dots, b_d + j_d(a_d - b_d)\big), \tag{17}$$

where $J_i := \mathbb{I}\{a_i \neq b_i\}$ for each i. For example, when $d = 2$, it is easy to see that $\Delta(f; [\mathbf{a}, \mathbf{b}])$ equals

$$\begin{aligned} f(b_1, b_2) - f(b_1, a_2) - f(a_1, b_2) + f(a_1, a_2) \quad &\text{if } \mathbf{a} \prec \mathbf{b} \\ f(b_1, b_2) - f(b_1, a_2) \quad &\text{if } a_1 = b_1 \text{ and } a_2 < b_2 \\ f(b_1, b_2) - f(a_1, b_2) \quad &\text{if } a_2 = b_2 \text{ and } a_1 < b_1. \end{aligned} \tag{18}$$

We are now ready to define entire monotonicity.

DEFINITION 2.1 (Entire monotonicity). We say that a function $f : [0, 1]^d \to \mathbb{R}$ is *entirely monotone* if

$$\Delta\big(f; [\mathbf{a}, \mathbf{b}]\big) \ge 0 \quad \text{for every } \mathbf{a} \neq \mathbf{b} \in [0, 1]^d \text{ with } \mathbf{a} \preceq \mathbf{b}.$$

In words, for a entirely monotone function f, every quasi-volume $\Delta(f; [\mathbf{a}, \mathbf{b}])$ is nonnegative. The class of such functions will be denoted by $\mathcal{F}_{\mathrm{EM}}^d$. By (18), note that entire monotonicity is equivalent to (3) for $d = 2$.

A more common generalization of monotonicity to multiple dimensions is the class $\mathcal{F}_{\mathrm{M}}^d$ consisting of all functions $f : [0, 1]^d \to \mathbb{R}$ satisfying

$$f(a_1, \dots, a_d) \le f(b_1, \dots, b_d), \quad \text{for } 0 \le a_i \le b_i \le 1, i = 1, \dots, d. \tag{19}$$

As the following result shows (see Section D.1 of the Supplementary Material [21] for a proof), $\mathcal{F}_{\mathrm{EM}}^d$ is a strict subset of $\mathcal{F}_{\mathrm{M}}^d$ when $d \ge 2$ (e.g., when $d = 2$, functions in $\mathcal{F}_{\mathrm{EM}}^d$ need to additionally satisfy the second constraint in (3)) and thus the estimator (2) is distinct from the LSE over $\mathcal{F}_{\mathrm{M}}^d$ for $d \ge 2$. This latter estimator is the classical multivariate isotonic regression estimator [46].

LEMMA 2.1. *When $d = 1$, entire monotonicity coincides with monotonicity, that is, $\mathcal{F}^1_{\mathrm{EM}} = \mathcal{F}^1_{\mathrm{M}}$. For $d \geq 2$, we have $\mathcal{F}^d_{\mathrm{EM}} \subsetneq \mathcal{F}^d_{\mathrm{M}}$.*

It is well known that entirely monotone functions are closely related to cumulative distribution functions of nonnegative measures. The following result taken from Aistleitner and Dick [1], Theorem 3, makes this connection precise.

LEMMA 2.2 ([1], Theorem 3).

1. *For every nonnegative Borel measure ν on $[0, 1]^d$, the function $f(\mathbf{x}) := \nu([\mathbf{0}, \mathbf{x}])$ belongs to $\mathcal{F}^d_{\mathrm{EM}}$.*
2. *If $f \in \mathcal{F}^d_{\mathrm{EM}}$ is right continuous, then there exists a unique nonnegative Borel measure ν on $[0, 1]^d$ such that $f(\mathbf{x}) - f(\mathbf{0}) = \nu([\mathbf{0}, \mathbf{x}])$.*

We shall now define the notion of HK**0** variation. The HK**0** variation is defined through another variation called the *Vitali variation*. Let us first define the Vitali variation of a function $f : [0, 1]^d \to \mathbb{R}$. To do so, we need some notation. By a partition of the univariate interval $[0, 1]$, we mean a set of points $0 = x_0 < x_1 < \cdots < x_k = 1$ for some $k \geq 1$. Given d such univariate partitions:

$$0 = x_0^{(s)} < x_1^{(s)} < \cdots < x_{k_s}^{(s)} = 1, \quad \text{for } s = 1, \ldots, d, \tag{20}$$

we can define a collection $\mathcal{P}$ of subsets of $[0, 1]^d$ consisting of all sets of the form $A_1 \times \cdots \times A_d$ where for each $1 \leq s \leq d$, $A_s = [x_{l_s}^{(s)}, x_{l_s+1}^{(s)}]$ for some $0 \leq l_s \leq k_s - 1$. Note that each set in $\mathcal{P}$ is an axis-aligned closed rectangle and the cardinality of $\mathcal{P}$ equals $k_1 \ldots k_d$. The rectangles in $\mathcal{P}$ are not disjoint but they form a *split* of $[0, 1]^d$ in the sense of Owen [44], Definition 3, and we shall refer to $\mathcal{P}$ as the split generated by the d univariate partitions (20).

DEFINITION 2.2 (Vitali variation). The Vitali variation of a function $f : [0, 1]^d \to \mathbb{R}$ is defined as

$$V^{(d)}(f; [0, 1]^d) := \sup_{\mathcal{P}} \sum_{A \in \mathcal{P}} |\Delta(f; A)|,$$

where $\Delta(f; A)$ is the quasi-volume defined in (17) and the supremum above is taken over all splits $\mathcal{P}$ that are generated by d univariate partitions in the manner described above.

The following observations about the Vitali variation will be useful for us. Note first that when $d = 1$, Vitali variation is simply total variation (4) since the rectangles in this case are intervals. The second fact is that when f is smooth (in the sense that the partial derivatives appearing below exist and are continuous on $[0, 1]^d$), we have

$$V^{(d)}(f; [0, 1]^d) = \int_0^1 \ldots \int_0^1 \left| \frac{\partial^d f}{\partial x_1 \ldots \partial x_d} \right| dx_1 \ldots dx_d. \tag{21}$$

The third observation is that $V^{(d)}(f; [0, 1]^d)$ can be written out explicitly when f is a rectangular piecewise constant function. In order to state this result, let us formally define the notion of a rectangular piecewise constant function on $[0, 1]^d$. Given d univariate partitions as in (20), let $\mathcal{P}^*$ denote the collection of all sets of the form $A_1 \times \cdots \times A_d$ where for each $1 \leq s \leq d$, A_s is either equal to $[x_{l_s}^{(s)}, x_{l_s+1}^{(s)})$ for some $0 \leq l_s \leq k_s - 1$ or the singleton $\{1\}$. Note that, unlike $\mathcal{P}$, the sets in $\mathcal{P}^*$ are disjoint, and hence $\mathcal{P}^*$ forms a partition of $[0, 1]^d$. We shall refer to $\mathcal{P}^*$ as the partition generated by the d univariate partitions (20).

DEFINITION 2.3 (Rectangular piecewise constant function). We say that $f : [0, 1]^d \to \mathbb{R}$ is rectangular piecewise constant if there exists a partition $\mathcal{P}^*$ generated by d univariate partitions as described above such that f is constant on each set in $\mathcal{P}^*$. We use $\mathfrak{R}^d$ to denote the class of all rectangular piecewise constant functions on $[0, 1]^d$. For $f \in \mathfrak{R}^d$, we define $k(f)$ as the smallest value of $k_1 \dots k_d$ for which there exist d univariate partitions of lengths $k_1, \dots, k_d$ such that f is constant on each of the sets in $\mathcal{P}^*$ generated by these d univariate partitions.

The following lemma (proved in Section D.2 of the Supplementary Material [21]) provides a formula for the Vitali variation of a rectangular piecewise constant function f on $[0, 1]^d$. Note that this lemma implies, in particular, that the Vitali variation of every rectangular piecewise constant function is finite.

LEMMA 2.3. *Suppose f is rectangular piecewise constant on $[0, 1]^d$ with respect to a partition $\mathcal{P}^*$ generated by d univariate partitions and let $\mathcal{P}$ denote the split generated by these univariate partitions. Then*

$$V^{(d)}(f; [0, 1]^d) = \sum_{A \in \mathcal{P}} |\Delta(f; A)|.$$

Despite these interesting properties, the Vitali variation is not directly suitable for our purposes because there exist many nonconstant functions f on $[0, 1]^d$ (such as $f(x, y) := x$) whose Vitali variation is zero. This weakness of the Vitali variation is well known (see, e.g., Owen [44] or Aistleitner and Dick [1]) and motivates the following definition of the HK**0** variation.

Given a nonempty subset of indices $S \subseteq [d] = \{1, \dots, d\}$, let

$$U_S := \{(u_1, \dots, u_d) \in [0, 1]^d : u_j = 0, j \notin S\}. \tag{22}$$

Note that U_S is a face of $[0, 1]^d$ adjacent to $\mathbf{0}$. By ignoring the components not in S, the restriction of the function f on $[0, 1]^d$ to the set U_S can be viewed as a function $\tilde{f} : [0, 1]^{|S|} \to \mathbb{R}$. The Vitali variation of $\tilde{f}$ viewed as a function of $[0, 1]^{|S|}$ will be denoted by

$$V^{(|S|)}(f; S; [0, 1]^d) := V^{(|S|)}(\tilde{f}; [0, 1]^{|S|}).$$

The *Hardy–Krause variation (anchored at* $\mathbf{0}$*)* of $f : [0, 1]^d \to \mathbb{R}$ is defined by

$$V_{\mathrm{HK}\mathbf{0}}(f; [0, 1]^d) := \sum_{\varnothing \neq S \subseteq [d]} V^{(|S|)}(f; S; [0, 1]^d).$$

That is, the HK**0** variation is the sum of the Vitali variations of f restricted to each face of $[0, 1]^d$ adjacent to $\mathbf{0}$. Note the special role played by the point $\mathbf{0}$ in this definition and this is the reason for the phrase "anchored at $\mathbf{0}$." It is also common to anchor the HK variation at $\mathbf{1}$ (see, e.g., Aistleitner and Dick [1]) but we focus only on $\mathbf{0}$ as the anchor in this paper. Because of the addition of the lower-dimensional Vitali variations, it is clear that the HK**0** variation equals zero only for constant functions and this property is the reason why the HK**0** variation is usually preferred to the Vitali variation.

Let us now remark that the HK**0** variation is quite different from the usual notion of multivariate total variation. Indeed, when f is smooth, the multivariate total variation of f only involves the first-order partial derivatives of f. On the other hand, as can be seen from (21), the HK**0** variation is defined in terms of higher order mixed partial derivatives of f.

An important property of the HK**0** variation is that it is finite for rectangular piecewise constant functions. This is basically a consequence of Lemma 2.3 and the fact that the restriction of a rectangular piecewise constant function to each set U_S in (22) is also rectangular piecewise constant.

The following lemma formally establishes the connection between entire monotonicity and HK**0** variation, as mentioned earlier in the Introduction.

LEMMA 2.4. *The following properties hold*:

(i) *If* $f : [0,1]^d \to \mathbb{R}$ *has finite* HK**0** *variation*, *then there exist unique* $f_+, f_- \in \mathcal{F}^d_{\mathrm{EM}}$ *such that* $f_+(\mathbf{0}) = f_-(\mathbf{0}) = 0$ *and*

$$f(\mathbf{x}) - f(\mathbf{0}) = f_+(\mathbf{x}) - f_-(\mathbf{x}), \quad \mathbf{x} \in [0,1]^d$$

and

$$V_{\mathrm{HK}\mathbf{0}}(f;[0,1]^d) = V_{\mathrm{HK}\mathbf{0}}(f_+;[0,1]^d) + V_{\mathrm{HK}\mathbf{0}}(f_-;[0,1]^d).$$

(ii) *If* $f \in \mathcal{F}^d_{\mathrm{EM}}$, *then*

$$V_{\mathrm{HK}\mathbf{0}}(f;[0,1]^d) = f(\mathbf{1}) - f(\mathbf{0}).$$

The first fact in the above lemma is quite standard (see, e.g., [1], Theorem 2). We could not find an exact reference for the second fact so we included a proof in Section D.3 of the Supplementary Material [21].

Finally, let us mention that it is well known that a result analogous to Lemma 2.2 holds for the connection between functions with finite HK**0** variation and cumulative distribution functions for signed measures. This result is stated next.

LEMMA 2.5 ([1], Theorem 3).

1. *For every signed Borel measure* ν *on* $[0,1]^d$, *the function* $f(\mathbf{x}) := \nu([\mathbf{0},\mathbf{x}])$ *has finite HK***0** *variation*.
2. *If* f *has finite* HK**0** *variation and is right continuous*, *then there exists a unique finite signed Borel measure* ν *on* $[0,1]^d$ *such that* $f(\mathbf{x}) = \nu([\mathbf{0},\mathbf{x}])$.

3. Computational feasibility.

The goal of this section is to describe procedures for computing the two estimators (2) and (6). We shall specifically show that the estimators (2) and (6) can be computed by solving a NNLS problem and a LASSO problem, respectively, with a suitable design matrix that is the same for both the problems and that depends only on $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This design matrix will be the matrix $\mathbf{A}$ whose columns are the distinct elements of the finite set

$$\mathcal{Q} \equiv \mathcal{Q}_{\mathbf{x}_1,\ldots,\mathbf{x}_n} := \{\mathbf{v}(\mathbf{z}) : \mathbf{z} \in [0,1]^d\} \subseteq \{0,1\}^n, \tag{23}$$

where

$$\mathbf{v}(\mathbf{z}) \equiv \mathbf{v}_{\mathbf{x}_1,\ldots,\mathbf{x}_n}(\mathbf{z}) := (\mathbb{I}_{[\mathbf{z},\mathbf{1}]}(\mathbf{x}_1), \mathbb{I}_{[\mathbf{z},\mathbf{1}]}(\mathbf{x}_2), \ldots, \mathbb{I}_{[\mathbf{z},\mathbf{1}]}(\mathbf{x}_n)).$$

We assume without loss of generality that the first column of $\mathbf{A}$ is $\mathbf{v}(\mathbf{0}) = \mathbf{1} = (1,\ldots,1) \in \mathbb{R}^n$. Note that $\mathbf{A}$ has dimensions $n \times p$ where $p \equiv p(\mathbf{x}_1,\ldots,\mathbf{x}_n) := |\mathcal{Q}|$. By definition, there exist distinct points $\mathbf{z}_1,\ldots,\mathbf{z}_p \in [0,1]^d$ with $\mathbf{z}_1 = \mathbf{0}$ such that the jth column of $\mathbf{A}$ is $\mathbf{v}(\mathbf{z}_j)$ for each j.

Our first result below deals with problem (2). Given the design matrix $\mathbf{A}$, we can define the following NNLS problem

$$\hat{\boldsymbol{\beta}}_{\mathrm{EM}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p:\beta_j\geq 0,\forall j\geq 2}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|^2, \tag{24}$$

where $\mathbf{y}$ is the $n \times 1$ vector consisting of the observations $y_1, \ldots, y_n$ coming from model (1). (24) is clearly a finite dimensional convex optimization problem (in fact, a quadratic optimization problem with linear constraints). Its solution $\widehat{\boldsymbol{\beta}}_{\mathrm{EM}}$ is not necessarily unique but the vector $\mathbf{A}\widehat{\boldsymbol{\beta}}_{\mathrm{EM}}$ is the projection of the observation vector $\mathbf{y}$ onto the closed convex cone $\{\mathbf{A}\boldsymbol{\beta} : \min_{j \geq 2} \beta_j \geq 0\}$ and is thus unique. The next result (proved in Section D.6 of the Supplementary Material [21]) shows how to obtain a solution to problem (2) using any solution $\widehat{\boldsymbol{\beta}}_{\mathrm{EM}}$ of (24).

PROPOSITION 3.1. *One solution for the optimization problem* (2) *is*

$$\hat{f}_{\mathrm{EM}} := \sum_{j=1}^{p} (\hat{\beta}_{\mathrm{EM}})_j \cdot \mathbb{I}_{[\mathbf{z}_j, \mathbf{1}]}, \tag{25}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{EM}} = ((\hat{\beta}_{\mathrm{EM}})_1, \ldots, (\hat{\beta}_{\mathrm{EM}})_p)$ *is any solution to* (24).

Thus, one way to compute the estimator (2) is to solve the NNLS problem (24) and use the resulting coefficients in the above manner (25). It is interesting to note that the solution (25) is a rectangular piecewise constant function and the quantity $k(\hat{f}_{\mathrm{EM}})$ (see Definition 2.3) will be controlled by the sparsity of $\widehat{\boldsymbol{\beta}}_{\mathrm{EM}}$. The key to proving Proposition 3.1 is the following characterization of $\mathcal{F}^d_{\mathrm{EM}}$ (proved in Section D.5 of the Supplementary Material [21]).

PROPOSITION 3.2 (Discretization of entirely monotone functions). *For every set of design points* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in [0,1]^d$, *we have*

$$\{\mathbf{A}\boldsymbol{\beta} : \beta_j \geq 0, \forall j \geq 2\} = \{(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) : f \in \mathcal{F}^d_{\mathrm{EM}}\}.$$

Note that Proposition 3.2 immediately implies that for every minimizer $\hat{f}_{\mathrm{EM}}$ of (2), the vector $(\hat{f}_{\mathrm{EM}}(\mathbf{x}_1), \ldots, \hat{f}_{\mathrm{EM}}(\mathbf{x}_n))$ equals $\mathbf{A}\widehat{\boldsymbol{\beta}}_{\mathrm{EM}}$ and is thus unique.

We now turn to problem (6). Given the matrix $\mathbf{A}$ and a tuning parameter $V > 0$, we can define the following LASSO problem:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{HK}\mathbf{0},V} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p : \sum_{j \geq 2} |\beta_j| \leq V}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|^2. \tag{26}$$

Again $\widehat{\boldsymbol{\beta}}_{\mathrm{HK}\mathbf{0},V}$ may not be unique but $A\widehat{\boldsymbol{\beta}}_{\mathrm{HK}\mathbf{0},V}$ is unique as it is the projection of $\mathbf{y}$ onto the closed convex set

$$\mathcal{C}(V) := \Big\{\mathbf{A}\boldsymbol{\beta} : \sum_{j \geq 2} |\beta_j| \leq V\Big\}. \tag{27}$$

The next result (proved in Section D.8 of the Supplementary Material [21]) shows how to obtain a solution to (6) using any solution $\widehat{\boldsymbol{\beta}}_{\mathrm{HK}\mathbf{0},V}$ of (26).

PROPOSITION 3.3. *One solution for the optimization problem* (6) *is*

$$\hat{f}_{\mathrm{HK}\mathbf{0},V} := \sum_{j=1}^{p} (\hat{\beta}_{\mathrm{HK}\mathbf{0},V})_j \cdot \mathbb{I}_{[\mathbf{z}_j, \mathbf{1}]}, \tag{28}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{HK}\mathbf{0},V} = ((\hat{\beta}_{\mathrm{HK}\mathbf{0},V})_1, \ldots, (\hat{\beta}_{\mathrm{HK}\mathbf{0},V})_p)$ *is the solution to the LASSO problem* (26).

Thus, one way to compute the estimator (6) is to solve the LASSO problem (26) and use the resulting coefficients to construct the rectangular piecewise constant function (6). Note the strong similarity between the two expressions (25) and (28). The following result (proved in Section D.7 of the Supplementary Material [21]) is the key ingredient in proving the above.

PROPOSITION 3.4. *For every set of design points* $\mathbf{x}_1, \dots, \mathbf{x}_n \in [0,1]^d$, *we have*

$$\mathcal{C}(V) = \{(f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)) : V_{\mathrm{HK}\mathbf{0}}(f; [0,1]^d) \le V\}.$$

Proposition 3.4 immediately implies that for every minimizer $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ of (6), the vector $(\widehat{f}_{\mathrm{HK}\mathbf{0},V}(\mathbf{x}_1), \dots, \widehat{f}_{\mathrm{HK}\mathbf{0},V}(\mathbf{x}_n))$ equals $\mathbf{A}\widehat{\boldsymbol{\beta}}_{\mathrm{HK}\mathbf{0},V}$ and is thus unique.

We have thus shown that the LSEs defined by (2) and (6) can be computed via NNLS and LASSO estimators with respect to the design matrix $\mathbf{A}$ whose columns are the elements of the finite set $\mathcal{Q}$ defined in (23). Once the design matrix $\mathbf{A}$ is formed, we can use existing quadratic program solvers to solve the NNLS and LASSO problems. The key to forming $\mathbf{A}$ is to enumerate the elements of $\mathcal{Q}$ and we address this issue now. We first state the following result which provides a worst case upper bound on $p \equiv p(\mathbf{x}_1, \dots, \mathbf{x}_n)$, the cardinality of $\mathcal{Q}$.

LEMMA 3.5. *The cardinality of* $\mathcal{Q}$ *satisfies*

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n) \le \sum_{j=0}^{d} \binom{n}{j} \tag{29}$$

for every $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$.

Lemma 3.5 is a consequence of the Vapnik–Chervonenkis lemma [58] and is proved in Section D.9 of the Supplementary Material [21]. Note that the upper bound (29) can be further bounded by $(en/d)^d$.

We emphasize here that Lemma 3.5 gives a worst case upper bound for $p(\mathbf{x}_1, \dots, \mathbf{x}_n)$ (here worst case is in terms of the design configurations $\mathbf{x}_1, \dots, \mathbf{x}_n$). For specific choices of $\mathbf{x}_1, \dots, \mathbf{x}_n$, the quantity $p(\mathbf{x}_1, \dots, \mathbf{x}_n)$ can be much smaller than the right-hand side of (29). For example, if $\mathbf{x}_1, \dots, \mathbf{x}_n$ are an enumeration of the grid points $\{(i_1/n^{1/d}, \dots, i_d/n^{1/d}) : i_1, \dots, i_d \in \{1, \dots, n^{1/d}\}\}$ (or form any other full grid), then $p(\mathbf{x}_1, \dots, \mathbf{x}_n) = n$ whereas the upper bound in (29) is of order n^d. However, there exist design configurations $\mathbf{x}_1, \dots, \mathbf{x}_n$ where the upper bound can be tight. For instance, when $d = 2$, if $\mathbf{x}_1, \dots, \mathbf{x}_n$ lie on the antidiagonal (the line segment connecting $(0,1)$ and $(1,0)$), then $p(\mathbf{x}_1, \dots, \mathbf{x}_n) = \frac{n(n+1)}{2}$, so the upper bound $\frac{n(n+1)}{2} + 1$ in (29) is nearly tight for $p(\mathbf{x}_1, \dots, \mathbf{x}_n)$.

The task of enumerating $\mathcal{Q}$ in general can be simplified if we show that we only need to check the value of $\mathbb{I}_{[\mathbf{z},\mathbf{1}]}$ on the design points $\mathbf{x}_1, \dots, \mathbf{x}_n$ for all $\mathbf{z}$ in some finite set S, rather than all $\mathbf{z} \in (0,1]^d$ as in definition (23). Then we can list all $|S|$ evaluation vectors (and remove duplicates if necessary) to form $\mathbf{A}$. The following two strategies can be used to construct the set S:

1. *Naïve gridding.* The simplest idea is to let S be the smallest grid that contains the design points $\mathbf{x}_1, \dots, \mathbf{x}_n$. That is, let $S = S_1 \times \cdots \times S_d$ where $S_j := \{(\mathbf{x}_1)_j, \dots, (\mathbf{x}_n)_j\}$ is the set of unique jth component values among the design points. It is simple to check that for any $\mathbf{z} \in (0,1]^d$, the value of $\mathbb{I}_{[\mathbf{z},\mathbf{1}]}$ on the design points is the same as $\mathbb{I}_{[\mathbf{z}',\mathbf{1}]}$, where $\mathbf{z}'$ is the smallest element of S such that $\mathbf{z} \preceq \mathbf{z}'$. In the worst case, $|S_j| = n$ for each j, so we would need to check at most $|S| = n^d$ vectors.
2. *Componentwise minimum.* A better approach is to let

$$S := \{\min\{\mathbf{x}_i : i \in I\} : I \subseteq [n], |I| \le d\},$$

where "min" denotes componentwise minimum of vectors. That is, for each subset of the design points of size $\le d$, we take the componentwise minimum and include that vector in S. To see why this definition of S suffices, consider any $\mathbf{z} \in [0,1]^d$ and note the $\mathbb{I}_{[\mathbf{z},\mathbf{1}]}$ has the same values on the design points as $\mathbb{I}_{[\mathbf{z}',\mathbf{1}]}$, where $\mathbf{z}' := \min\{\mathbf{x}_i : i \in J\}$ and $J := \{i : \mathbf{z} \preceq \mathbf{x}_i\}$.

Furthermore, by the same reasoning as in our VC dimension computation above, there must exist some subset $I \subseteq J$ of size $\le d$ such that $\min\{\mathbf{x}_i : i \in J\} = \min\{\mathbf{x}_i : i \in I\}$, which proves $\mathbf{z}' \in S$. In the worst case, we would need to check $|S| = \sum_{j=0}^{d} \binom{n}{j}$ vectors, which is the VC upper bound (29).

3.1. *Special case*: *The equally-spaced lattice design.* The results stated so far in the section hold for every configuration of design points $\mathbf{x}_1, \dots, \mathbf{x}_n \in [0, 1]^d$. We now specialize to the setting where $\mathbf{x}_1, \dots, \mathbf{x}_n$ form an equally-spaced lattice (precisely defined below). Our theoretical results described in the next section work under this setting. Moreover, some of the estimators from the literature that are related to $\widehat{f}_{\mathrm{EM}}$ and $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ are defined only under the lattice design so a discussion of the form of our estimators in this setting will make it easier for us to compare and contrast them with existing estimators (this comparison is the subject of Section 5).

Given positive integers $n_1, \dots, n_d$ with $n = n_1 \dots n_d$, by a lattice design of dimensions $n_1 \times \dots \times n_d$, we mean that $\mathbf{x}_1, \dots, \mathbf{x}_n$ form an enumeration of the points in

$$\mathbb{L}_{n_1,\dots,n_d} := \{(i_1/n_1, \dots, i_d/n_d) : 0 \le i_j \le n_j - 1, j = 1, \dots, d\} \tag{30}$$

Note that, in this setting, the set $\mathcal{Q}$ (defined in (23)) can be enumerated by $\mathcal{Q} = \{\mathbf{v}(\mathbf{x}_1), \dots, \mathbf{v}(\mathbf{x}_n), \mathbf{0}\}$. Without loss of generality, we may ignore the $\mathbf{0}$ element and assume the columns of $\mathbf{A}$ are $\mathbf{v}(\mathbf{x}_1), \dots, \mathbf{v}(\mathbf{x}_n)$ so that the i, j entry of $\mathbf{A}$ is given by $\mathbf{A}(i, j) = \mathbb{I}_{[\mathbf{x}_j, \mathbf{1}]}(\mathbf{x}_i) = \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\}$. We also take $\mathbf{x}_1 := \mathbf{0}$ (corresponding to $i_1 = \dots = i_d = 0$) so that the first column of $\mathbf{A}$ is the vector of ones. Therefore, in the lattice design setting, the optimization problems (24) and (26) for computing the two estimators $\widehat{f}_{\mathrm{EM}}$ and $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ can be rewritten as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{EM}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j \ge 0, \forall j \ge 2}{\operatorname{argmin}} \sum_{i=1}^{n} \Bigg(y_i - \sum_{j=1}^{n} \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\} \boldsymbol{\beta}_j \Bigg)^2 \tag{31}$$

and

$$\widehat{\boldsymbol{\beta}}_{\mathrm{HK}\mathbf{0},V} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p : \sum_{j \ge 2} |\beta_j| \le V}{\operatorname{argmin}} \sum_{i=1}^{n} \Bigg(y_i - \sum_{j=1}^{n} \mathbb{I}\{\mathbf{x}_j \preceq \mathbf{x}_i\} \boldsymbol{\beta}_j \Bigg)^2, \tag{32}$$

respectively. It also turns out that, in the lattice design setting, the matrix $\mathbf{A}$ is square and invertible (Lemma D.1). As a result, it is possible to write down the vectors $(\widehat{f}_{\mathrm{EM}}(\mathbf{x}_1), \dots, \widehat{f}_{\mathrm{EM}}(\mathbf{x}_n))$ and $(\widehat{f}_{\mathrm{HK}\mathbf{0},V}(\mathbf{x}_1), \dots, \widehat{f}_{\mathrm{HK}\mathbf{0},V}(\mathbf{x}_n))$ as solutions to more explicit constrained quadratic optimization problems. This is the content of the next result which is proved in Section D.10 of the Supplementary Material [21]. Here, it will be convenient to represent vectors in $\mathbb{R}^n$ as tensors indexed by $\mathbf{i} := (i_1, \dots, i_d) \in \mathcal{I}$ where

$$\mathcal{I} := \{\mathbf{i} = (i_1, \dots, i_d) : i_j \in \{0, 1, \dots, n_j - 1\} \text{ for every } j = 1, \dots, d\}.$$

In other words, we write the components of a vector $\boldsymbol{\theta} \in \mathbb{R}^n$ by $\theta_{\mathbf{i}}$ for $\mathbf{i} = (i_1, \dots, i_d) \in \mathcal{I}$. We will also denote the observation corresponding to the design point $(i_1/n_1, \dots, i_d/n_d)$ by $y_{\mathbf{i}} = y_{i_1,\dots,i_d}$.

LEMMA 3.6. *Consider the setting of the lattice design of dimensions $n_1 \times \dots \times n_d$. For each $\boldsymbol{\theta} \in \mathbb{R}^n$, associate the "differenced" vector $D\boldsymbol{\theta} \in \mathbb{R}^n$ whose $\mathbf{i}^{th}$ entry is given by*

$$\sum_{j_1=0}^{1} \cdots \sum_{j_d=0}^{1} I\{i_1 - j_1 \ge 0, \dots, i_d - j_d \ge 0\} (-1)^{j_1+\dots+j_d} \theta_{i_1-j_1,\dots,i_d-j_d}$$

for every $\mathbf{i} = (i_1, \dots, i_d) \in \mathcal{I}$. Then:

1. *The vector* $(\widehat{f}_{\mathrm{EM}}(i_1/n_1,\dots,i_d/n_d):\mathbf{i}=(i_1,\dots,i_d)\in\mathcal{I})$ *is the solution to the optimization problem*

$$\operatorname{argmin}\Big\{\sum_{\mathbf{i}\in\mathcal{I}}(y_{\mathbf{i}}-\theta_{\mathbf{i}})^2:(D\boldsymbol{\theta})_{\mathbf{i}}\geq 0\, \textit{for all}\, \mathbf{i}\neq\mathbf{0}\Big\}. \tag{33}$$

2. *The vector* $(\widehat{f}_{\mathrm{HK}\mathbf{0},V}(i_1/n_1,\dots,i_d/n_d):\mathbf{i}=(i_1,\dots,i_d)\in\mathcal{I})$ *is the solution to the optimization problem*

$$\operatorname{argmin}\Big\{\sum_{\mathbf{i}}(y_{\mathbf{i}}-\theta_{\mathbf{i}})^2:\sum_{\mathbf{i}\neq\mathbf{0}}\big|(D\boldsymbol{\theta})_{\mathbf{i}}\big|\leq V\Big\}. \tag{34}$$

REMARK 3.1 (The special case of $d=2$). When $d=2$, it is easy to see that the differenced vector $D\boldsymbol{\theta}$ is given by

$$(D\boldsymbol{\theta})_{(i_1,i_2)}=\begin{cases}\theta_{i_1,i_2}-\theta_{i_1-1,i_2}-\theta_{i_1,i_2-1}+\theta_{i_1-1,i_2-1} & \text{if } i_1>0, i_2>0,\\ \theta_{i_1,0}-\theta_{i_1-1,0} & \text{if } i_1>0, i_2=0,\\ \theta_{0,i_2}-\theta_{0,i_2-1} & \text{if } i_1=0, i_2>0,\\ \theta_{0,0} & \text{if } i_1=i_2=0.\end{cases}$$

Using this, it is easy to see that (34) can be rewritten for $d=2$ as

$$\begin{aligned}\operatorname{argmin}\Big\{&\sum_{i_1=0}^{n_1-1}\sum_{i_2=0}^{n_2-1}(y_{i_1,i_2}-\theta_{i_1,i_2})^2:\\ &\sum_{i_1=1}^{n_1-1}\sum_{i_2=1}^{n_2-1}|\theta_{i_1,i_2}-\theta_{i_1-1,i_2}-\theta_{i_1,i_2-1}+\theta_{i_1-1,i_2-1}|\\ &+\sum_{i_1=1}^{n_1-1}|\theta_{i_1,0}-\theta_{i_1-1,0}|+\sum_{i_2=1}^{n_2-1}|\theta_{0,i_2}-\theta_{0,i_2-1}|\leq V\Big\}\end{aligned} \tag{35}$$

and a similar formula can be written for (33) for $d=2$.

As mentioned in the Introduction, an estimator similar to $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ has been described by Mammen and van de Geer [37] for $d=2$ under the lattice design setting. Specifically, the estimator of [37] for the vector $(f^*(i_1/n_1,i_2/n_2), 0\leq i_1\leq n_1-1, 0\leq i_2\leq n_2-1)$ is given by the solution to the optimization problem:

$$\begin{aligned}\operatorname{argmin}\Big\{&\sum_{i_1,i_2}(y_{i_1,i_2}-\theta_{i_1,i_2})^2\\ &+\lambda_1\sum_{i_1,i_2\geq 1}|\theta_{i_1,i_2}-\theta_{i_1-1,i_2}-\theta_{i_1,i_2-1}+\theta_{i_1-1,i_2-1}|\\ &+\lambda_2\sum_{i_1\geq 1}\big|\overline{\theta}^{(1)}_{i_1}-\overline{\theta}^{(1)}_{i_1-1}\big|+\lambda_2\sum_{i_2\geq 1}\big|\overline{\theta}^{(2)}_{i_2}-\overline{\theta}^{(2)}_{i_2-1}\big|\Big\},\end{aligned} \tag{36}$$

where λ_1 and λ_2 are positive tuning parameters, $\overline{\theta}^{(1)}_{i_1}:=\frac{1}{n_2}\sum_{i_2=0}^{n_2-1}\theta_{i_1,i_2}$ and $\overline{\theta}^{(2)}_{i_2}:=\frac{1}{n_1}\sum_{i_1=0}^{n_1-1}\theta_{i_1,i_2}$. This optimization problem is similar to (35) in that the first term in the penalty is the same in both problems. However the remaining terms in the penalty above are different from the terms in (35) although they are of the same spirit in that both are penalizing lower dimensional variations. Moreover, our estimator (35) has one tuning parameter (in the constrained form) and (36) has two tuning parameters in the penalized form. It should also be noted that we defined our estimators for arbitrary design points $\mathbf{x}_1, \dots, \mathbf{x}_n$ while Mammen and van de Geer [37] only considered the lattice design for $d = 2$.

4. Risk results. In this section, risk bounds for the estimators $\widehat{f}_{\mathrm{EM}}$ and $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ are presented. We define risk under the standard fixed design squared error loss function (see (9)). Throughout this section, we assume that we are working with the lattice design of dimensions $n_1 \times \cdots \times n_d$ with $n = n_1 \times \cdots \times n_d$ and $n_j \geq 1$ for all $j = 1, \dots, d$.

4.1. *Risk results for* $\widehat{f}_{\mathrm{EM}}$. In this subsection, we present bounds on the risk $\mathcal{R}(\widehat{f}_{\mathrm{EM}}, f^*)$ of $\widehat{f}_{\mathrm{EM}}$ under the well-specified assumption where we assume that $f^* \in \mathcal{F}^d_{\mathrm{EM}}$. The first result below (proved in Section C.2 of the Supplementary Material [21]) bounds the risk in terms of the HK$\mathbf{0}$ variation of f^*. Note that from part (ii) of Lemma 2.4, $V_{\mathrm{HK}\mathbf{0}}(f^*; [0,1]^d) = f^*(\mathbf{1}) - f^*(\mathbf{0})$ as $f^* \in \mathcal{F}^d_{\mathrm{EM}}$.

THEOREM 4.1. *Let* $f^* \in \mathcal{F}^d_{\mathrm{EM}}$ *and* $V^* := V_{\mathrm{HK}\mathbf{0}}(f^*; [0,1]^d)$. *For the lattice design* (30), *the estimator* $\widehat{f}_{\mathrm{EM}}$ *satisfies*

$$\begin{aligned}\mathcal{R}(\widehat{f}_{\mathrm{EM}}, f^*) \leq{} & C_d\left(\frac{\sigma^2 V^*}{n}\right)^{\frac{2}{3}}\left(\log\left(2 + \frac{V^*\sqrt{n}}{\sigma}\right)\right)^{\frac{2d-1}{3}} \\ & + C_d\frac{\sigma^2}{n}(\log(en))^{\frac{3d}{2}}(\log(e\log(en)))^{\frac{2d-1}{2}},\end{aligned} \tag{37}$$

where C_d *is a constant that depends only on the dimension* d.

Note that the bound (10) in the Introduction is the dominant first term of this bound (37).

REMARK 4.1 (Model misspecification). Theorem 4.1 is stated under the well-specified assumption $f^* \in \mathcal{F}^d_{\mathrm{EM}}$. In the misspecified setting where $f^* \notin \mathcal{F}^d_{\mathrm{EM}}$, our LSE $\widehat{f}_{\mathrm{EM}}$ will not be close to f^*, but rather to

$$\widetilde{f} \in \operatorname*{argmin}_{f \in \mathcal{F}^d_{\mathrm{EM}}} \sum_{i=1}^n \bigl(f(\mathbf{x}_i) - f^*(\mathbf{x}_i)\bigr)^2,$$

so it is reasonable to consider $\mathcal{R}(\widehat{f}_{\mathrm{EM}}, \widetilde{f})$ rather than $\mathcal{R}(\widehat{f}_{\mathrm{EM}}, f^*)$. By the argument outlined in Remark C.1, one can show that $\mathcal{R}(\widehat{f}_{\mathrm{EM}}, \widetilde{f})$ is upper bounded by the right-hand side of (37) after redefining V^* as $V_{\mathrm{HK}\mathbf{0}}(\widetilde{f}; [0,1]^d)$.

As mentioned in the Introduction, when $d = 1$, the estimator $\widehat{f}_{\mathrm{EM}}$ is simply the isotonic LSE for which Zhang [63] proved that

$$\mathcal{R}(\widehat{f}_{\mathrm{EM}}, f^*) \leq C\left(\frac{\sigma^2 V^*}{n}\right)^{\frac{2}{3}} + C\frac{\sigma^2}{n}\log(en) \tag{38}$$

for some constant $C > 0$. It is interesting to note that our risk bound (37) for general $d \geq 2$ has the same terms as the univariate bound (38) with additional logarithmic factors which depend on d. It is natural to ask therefore if these additional logarithmic factors are indeed necessary or merely artifacts of our analysis. The next result (a minimax lower bound) shows that every estimator pays a logarithmic multiplicative price of $\log n$ for $d = 2$ and $(\log n)^{2(d-2)/3}$ for $d \geq 3$ in the first $n^{-2/3}$ term. We do not, unfortunately, know if the $(\log n)^{3d/2}(\log\log n)^{(2d-1)/2}$

factor in the second term in (37) is necessary or artifactual, although we can prove that it can be removed by a modification of the estimator $\widehat{f}_{\mathrm{EM}}$ (see Theorem 4.3 below).

The next result (proved in Section C.7 of the Supplementary Material [21]) proves a lower bound for the minimax risk:

$$\mathfrak{M}_{\mathrm{EM},\sigma,V,d}(n) := \inf_{\widehat{f}_n} \sup_{f^* \in \mathcal{F}^d_{\mathrm{EM}}: V_{\mathrm{HK}\mathbf{0}}(f^*) \le V} \mathbb{E}_{f^*} \mathcal{L}(\widehat{f}_n, f^*), \tag{39}$$

where the expectation is with respect to model (1).

THEOREM 4.2. *Let $d \ge 2$, $V > 0$, $\sigma > 0$ and let $n_j \ge c_s n^{1/d}$ for all $j = 1, \dots, d$ for some $c_s \in (0, 1]$. Then there exists a positive constant C_d depending only on d and c_s, such that the minimax risk on the lattice design* (30) *satisfies*

$$\mathfrak{M}_{\mathrm{EM},\sigma,V,d}(n) \ge C_d \left(\frac{\sigma^2 V}{n}\right)^{\frac{2}{3}} \left(\log\left(\frac{V\sqrt{n}}{\sigma}\right)\right)^{\frac{2(d-2)}{3}}$$

provided n is larger than a positive constant $c_{d,\sigma^2/V^2}$ depending only on d, σ^2/V^2, and c_s. In the case $d = 2$, this bound can be tightened to

$$\mathfrak{M}_{\mathrm{EM},\sigma,V,d}(n) \ge C\left(\frac{\sigma^2 V}{n}\right)^{\frac{2}{3}} \log\left(\frac{V\sqrt{n}}{\sigma}\right). \tag{40}$$

Note that the assumption $n_j \ge c_s n^{1/d}$ for all j is reasonable, since if, for instance, $n_{d'+1} = n_{d'+2} \cdots = n_d = 1$ then we simply have a d'-dimensional problem where $d' < d$, which should have a smaller minimax risk.

As mentioned before, the above result shows that some dependence on dimension d in the logarithmic term cannot be avoided for any estimator. Note also, that for $d = 2$, the minimax lower bound (40) matches our upper bound in Theorem 4.1 implying minimaxity of $\widehat{f}_{\mathrm{EM}}$ for $d = 2$. For $d > 2$, there remains a gap of $\log n$ between our minimax lower bound and the upper bound in Theorem 4.1. This gap is due to a logarithmic gap between an upper bound and lower bound given by Blei et al. [7], Theorem 1.1, for the metric entropy of cumulative distribution functions of probability measures on $[0, 1]^d$, a gap that essentially reduces to improving estimates of a small ball probability of Brownian sheets (see discussion in [7] for more detail and references).

As mentioned earlier, the logarithmic factor $(\log n)^{3d/2}(\log\log n)^{(2d-1)/2}$ appearing in the second term of (37) can be removed by a modification of the estimator $\widehat{f}_{\mathrm{EM}}$. This is shown in the next result. For a tuning parameter $V \ge 0$, let

$$\widetilde{f}_{\mathrm{EM},V} \in \operatorname*{argmin}_{f \in \mathcal{F}^d_{\mathrm{EM}}: V_{\mathrm{HK}\mathbf{0}}(f) \le V} \frac{1}{n}\sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2.$$

Note that this differs from the original estimator (2) only by the introduction of the additional constraint $V_{\mathrm{HK}\mathbf{0}}(f) \le V$.

THEOREM 4.3. *Let $f^* \in \mathcal{F}^d_{\mathrm{EM}}$ and $V^* := V_{\mathrm{HK}\mathbf{0}}(f^*; [0, 1]^d)$. Assume the lattice design* (30). *If the tuning parameter V is such that $V \ge V^*$, then the estimator $\widetilde{f}_{\mathrm{EM},V}$ satisfies*

$$\mathcal{R}(\widetilde{f}_{\mathrm{EM},V}, f^*) \le C_d \left(\frac{\sigma^2 V}{n}\right)^{\frac{2}{3}} \left(\log\left(2 + \frac{V\sqrt{n}}{\sigma}\right)\right)^{\frac{2d-1}{3}} + C_d \frac{\sigma^2}{n}.$$

Note that the second term in (4.3) is just σ^2/n and smaller than the second term in (37) but this comes at the cost of introducing a tuning parameter V that needs to be at least V^*.

We will now prove near-parametric rates for $\widehat{f}_{\mathrm{EM}}$ when f^* is rectangular piecewise constant. To motivate these results, note first that when f^* is constant on $[0,1]^d$, we have $V^* = 0$, and thus the bound given by (37) is σ^2/n up to logarithmic factors. In the next result (proved in Section C.3 of the Supplementary Material [21]), we generalize this fact and show that $\widehat{f}_{\mathrm{EM}}$ achieves nearly the parametric rate for rectangular piecewise constant functions $f^* \in \mathcal{F}^d_{\mathrm{EM}}$. Recall the definition of the class $\mathfrak{R}^d$ of all rectangular piecewise constant functions and the associated mapping $k(f)$, $f \in \mathfrak{R}^d$, from Definition 2.3.

THEOREM 4.4. *For every* $f^* : [0,1]^d \to \mathbb{R}$, *the LSE* $\widehat{f}_{\mathrm{EM}}$ *satisfies*

$$\mathcal{R}(\widehat{f}_{\mathrm{EM}}, f^*) \le \inf_{f \in \mathfrak{R}^d \cap \mathcal{F}^d_{\mathrm{EM}}} \left\{ \mathcal{L}(f, f^*) + C_d \sigma^2 \frac{k(f)}{n} (\log(en))^{\frac{3d}{2}} (\log(e \log(en)))^{\frac{2d-1}{2}} \right\}.$$

Theorem 4.4 gives a sharp oracle inequality in the sense of [5] as it applies to every function f^* (even in the misspecified case when $f^* \notin \mathcal{F}^d_{\mathrm{EM}}$) and the constant in front of the first term inside the infimum equals 1. Even though the inequality holds for every f^*, the right-hand side will be small only when f^* is close to some function f in $\mathfrak{R}^d \cap \mathcal{F}^d_{\mathrm{EM}}$. This implies that when $f^* \in \mathfrak{R}^d \cap \mathcal{F}^d_{\mathrm{EM}}$, we can take $f = f^*$ in the right-hand side to obtain that the risk of $\widehat{f}_{\mathrm{EM}}$ decays as $\sigma^2 k(f^*)/n$ up to logarithmic factors. This rate will be faster than the rate given by Theorem 4.1 provided $k(f^*)$ is not too large. Note that one can combine the two bounds given by Theorem 4.1 and Theorem 4.4 by taking their minimum. In the case $d = 1$, Theorem 4.4 reduces to the adaptive rates for isotonic regression [5, 12] but with worse logarithmic factors.

We would also like to mention here that $\mathfrak{R}^d \cap \mathcal{F}^d_{\mathrm{EM}}$ is a smaller class compared to $\mathfrak{R}^d \cap \mathcal{F}^d_{\mathrm{M}}$ (recall that $\mathcal{F}^d_{\mathrm{M}}$ is defined via (19)). Risk results over the class $\mathfrak{R}^d \cap \mathcal{F}^d_{\mathrm{M}}$ for the LSE over $\mathcal{F}^d_{\mathrm{M}}$ and other related estimators have been proved in Han et al. [30] and Deng and Zhang [17].

Before closing this subsection, let us briefly describe the main ideas underlying the proofs of Theorems 4.1, 4.2, 4.3 and 4.4. For Theorem 4.1, we use standard results on the accuracy of LSEs on closed convex sets which related the risk of $\widehat{f}_{\mathrm{EM}}$ to covering numbers of local balls of the form $\{f \in \mathcal{F}^d_{\mathrm{EM}} : \mathcal{L}(f, f^*) \le t^2\}$ for $t > 0$ sufficiently small in the pseudometric given by the square-root of the loss function $\mathcal{L}$. We calculated the covering numbers of these local balls by relating the functions in $\mathcal{F}^d_{\mathrm{EM}}$ to distribution functions of signed measures on $[0,1]^d$ and using existing covering number results for distribution functions of signed measures from Blei et al. [7] and Gao [23]. The proof of Theorem 4.2 is also based on covering number arguments as we use general minimax lower bounds from Yang and Barrons [61]. Finding lower bounds for the covering numbers under the pseudometric $\sqrt{\mathcal{L}}$ seems somewhat involved and we used a multiscale construction from Blei et al. [7], Section 4, for this purpose. The bound in Theorem 4.3 for $\widetilde{f}_{\mathrm{EM},V}$ is a quick consequence of the proof of the risk bound for $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ (Theorem 4.5) which is stated in the next subsection. For Theorem 4.4, we used standard results relating $\mathcal{R}(\widehat{f}_{\mathrm{EM}}, f^*)$ to a certain size-related measure (statistical dimension) of the tangent cone to $\widehat{f}_{\mathrm{EM}}$ at f^*. When $f^* \in \mathfrak{R}^d$ (or when f^* is approximable by a function in $\mathfrak{R}^d$), this tangent cone is decomposable into tangent cones of certain lower-dimensional tangent cones. The statistical dimension of these lower-dimensional tangent cones is then bounded via an application of Theorem 4.1 in the case when $V^* = 0$.

4.2. *Risk results for* $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$. In this subsection, we present bounds on the risk $\mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, f^*)$ of the estimator $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$. Note that the estimator $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ involves a tuning parameter V and therefore these results will require some conditions on V. Our first result below assumes that $V \ge V^* := V_{\mathrm{HK}\mathbf{0}}(f^*; [0,1]^d)$ and gives the $n^{-2/3}$ rate up to logarithmic factors. The proof of this result is given in Section C.4.

THEOREM 4.5. *Assume the lattice design* (30). *If the tuning parameter V is such that $V \geq V^* := V_{\mathrm{HK}\mathbf{0}}(f^*; [0,1]^d)$, then the estimator $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ satisfies*

$$\mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, f^*) \leq C_d \left(\frac{\sigma^2 V}{n}\right)^{\frac{2}{3}} \left(\log\left(2 + \frac{V\sqrt{n}}{\sigma}\right)\right)^{\frac{2d-1}{3}} + C_d \frac{\sigma^2}{n}. \tag{41}$$

REMARK 4.2. As mentioned earlier, Mammen and van de Geer [37] (see also the very recent paper Ortelli and van de Geer [42]) proposed the estimator (36) that is similar to $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$. Mammen and van de Geer [37] also proved a risk result for their estimator giving the rate $n^{-(1+d)/(1+2d)}$ which is strictly suboptimal compared to our rate in (41) for $d \geq 2$. This suboptimality is likely due to the use of suboptimal covering number bounds in [37].

REMARK 4.3 (Model misspecification). Theorem 4.5 is stated under the well-specified assumption $V_{\mathrm{HK}\mathbf{0}}(f^*; [0,1]^d) \leq V$. In the misspecified setting where $V_{\mathrm{HK}\mathbf{0}}(f^*; [0,1]^d) > V$, our LSE $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ will not be close to f^*, but to $\tilde{f} \in \operatorname{argmin}_{f: V_{\mathrm{HK}\mathbf{0}}(f) \leq V} \sum_{i=1}^n (f(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2$, so it is reasonable to consider $\mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, \tilde{f})$ rather than $\mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, f^*)$. By the argument outlined in Remark C.1 (in the Supplementary Material [21]), $\mathcal{R}(\widehat{f}_{\mathrm{EM}}, \tilde{f})$ is upper bounded by the right-hand side of (41).

In the next result, we prove a complementary minimax lower bound to Theorem 4.5 which proves that, for $d \geq 2$, the risk of every estimator over the class $\{f^* : V_{\mathrm{HK}\mathbf{0}}(f^*) \leq V\}$ is bounded from below by $n^{-2/3}(\log n)^{2(d-1)/3}$ (ignoring terms depending on d, V and σ). This implies that the logarithmic terms in (41) can perhaps be reduced slightly but cannot be removed altogether and must necessarily increase with the dimension d. Let

$$\mathfrak{M}_{\mathrm{HK},\sigma,V,d}(n) := \inf_{\widehat{f}_n} \sup_{f^*: V_{\mathrm{HK}\mathbf{0}}(f^*) \leq V} \mathbb{E}_{f^*} \mathcal{L}(\widehat{f}_n, f^*),$$

where the expectation is with respect to model (1). Note that $\{f^* \in \mathcal{F}^d_{\mathrm{EM}} : V_{\mathrm{HK}\mathbf{0}}(f^*) \leq V\} \subseteq \{f^* : V_{\mathrm{HK}\mathbf{0}}(f^*) \leq V\}$ which implies that

$$\mathfrak{M}_{\mathrm{HK},\sigma,V,d}(n) \geq \mathfrak{M}_{\mathrm{EM},\sigma,V,d}(n),$$

where $\mathfrak{M}_{\mathrm{EM},\sigma,V,d}(n)$ is defined in (39). This implies, in particular, that the lower bounds on $\mathfrak{M}_{\mathrm{EM},\sigma,V,d}(n)$ from Theorem 4.2 are also lower bounds on $\mathfrak{M}_{\mathrm{HK},\sigma,V,d}(n)$. However, the next result (whose proof is in Section C.6 of the Supplementary Material [21]) gives a strictly larger lower bound for $\mathfrak{M}_{\mathrm{HK},\sigma,V,d}(n)$ for $d > 2$ than that given by Theorem 4.2.

THEOREM 4.6. *Let $d \geq 2$, $V > 0$, $\sigma > 0$ and let $n_j \geq c_s n^{1/d}$ for $j = 1, \ldots, d$, where $c_s \in (0,1]$. Then there exists a positive constant C_d depending only on d and c_s, such that*

$$\mathfrak{M}_{\mathrm{HK},\sigma,V,d}(n) \geq C_d \left(\frac{\sigma^2 V}{n}\right)^{\frac{2}{3}} \left(\log\left(\frac{V\sqrt{n}}{\sigma}\right)\right)^{\frac{2(d-1)}{3}}$$

provided n is larger than a positive constant $c_{d,\sigma^2/V^2}$ depending only on d, σ^2/V^2, and c_s. In the case $d = 2$, this bound can be tightened to

$$\mathfrak{M}_{\mathrm{HK},\sigma,V,d}(n) \geq C \left(\frac{\sigma^2 V}{n}\right)^{\frac{2}{3}} \log\left(\frac{V\sqrt{n}}{\sigma}\right).$$

Theorems 4.5 and 4.6 together imply that $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ is minimax optimal over $\{f^* : V_{\mathrm{HK}\mathbf{0}}(f^*) \leq V\}$ for $d = 2$ and only possibly off by a factor of $(\log n)^{1/3}$ for $d > 2$.

We next explore the possibility of near parametric rates for $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ for rectangular piecewise constant functions. In the univariate case $d = 1$, it is known (see [28], Theorem 2.2) that $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ satisfies the near-parametric risk bound (13) provided (a) the tuning parameter V is taken to be close to V^*, (b) f^* is piecewise constant, and (c) the length of each constant piece of f^* is bounded from below by $c/k(f^*)$ for some $c > 0$. The next result (proved in Section C.8 of the Supplementary Material [21]) provides evidence that a similar story holds true for estimating certain rectangular piecewise constant functions.

For a given constant $0 < c \le 1/2$, let $\mathfrak{R}_1^d(c)$ denote the collection of functions $f : [0,1]^d \to \mathbb{R}$ of the form

$$f = a_1 \mathbb{I}_{[\mathbf{x}^*,\mathbf{1}]} + a_0 \tag{42}$$

for some $a_1, a_0 \in \mathbb{R}$ and $\mathbf{x}^* \in [0,1]^d$ satisfying the minimum size condition

$$\min\{|\mathbb{L}_{n_1,\dots,n_d} \cap [\mathbf{x}^*,\mathbf{1}]|, |\mathbb{L}_{n_1,\dots,n_d} \cap [\mathbf{0},\mathbf{x}^*)|\} \ge cn. \tag{43}$$

To gain more intuition about the above condition, note first that we are working with the lattice design so that $\mathbb{L}_{n_1,\dots,n_d} = \{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ is the set containing all design points. Roughly speaking, (43) ensures that $\mathbf{x}^*$ is not too close to the boundary of $[0,1]^d$ so that each of the rectangles $[\mathbf{x}^*,\mathbf{1}]$ and $[\mathbf{0},\mathbf{x}^*)$ contain at least some constant fraction of the n design points.

It is clear that $\mathfrak{R}_1^d(c)$ is a subset of $\mathfrak{R}^d$, that is, every function of the form (42) is rectangular piecewise constant. Indeed, it is easy to see that $k(f) \le 2^d$ for every $f \in \mathfrak{R}_1^d(c)$. The following result (proved in Section C.8 of the Supplementary Material [21]) bounds the risk of $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ for $f^* \in \mathfrak{R}_1^d(c)$.

THEOREM 4.7. *Consider the lattice design* (30) *with* $n > 1$. *Fix* $f^* : [0,1]^d \to \mathbb{R}$ *and consider the estimator* $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ *with a tuning parameter* V. *Then for every* $0 < c \le 1/2$, *we have*

$$\mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, f^*) \le \inf_{\substack{f \in \mathfrak{R}_1^d(c):\\ V_{\mathrm{HK}\mathbf{0}}(f)=V}} \left\{\mathcal{L}(f,f^*) + C(c,d)\frac{\sigma^2}{n}(\log n)^{\frac{3d}{2}}(\log\log n)^{\frac{2d-1}{2}}\right\} \tag{44}$$

for a constant $C(c,d)$ *that depends only on* c *and* d.

Theorem 4.7 applies to every function f^* but the infimum on the right-hand side of (44) is over all functions f in $\mathfrak{R}_1^d(c)$ with $V_{\mathrm{HK}\mathbf{0}}(f) = V$. Therefore, Theorem 4.7 implies that the risk of the estimator $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ with tuning parameter V at f^* is the near-parametric rate $\frac{\sigma^2}{n}(\log en)^{3d/2}(\log\log n)^{(2d-1)/2}$ provided f^* is close to some function f in $\mathfrak{R}_1^d(c)$ with $V = V_{\mathrm{HK}\mathbf{0}}(f)$. As an immediate consequence, we obtain that if $f^* \in \mathfrak{R}_1^d(c)$ and $V = V_{\mathrm{HK}\mathbf{0}}(f^*)$, then

$$\mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, f^*) \le C(c,d)\frac{\sigma^2}{n}\big(\log(en)\big)^{\frac{3d}{2}}\big(\log(e\log(en))\big)^{\frac{2d-1}{2}}.$$

Functions in $\mathfrak{R}_1^d(c)$ are constrained to satisfy the minimum size condition (43). A comparison of Theorem 4.7 with the corresponding univariate results shows that the near-parametric rate cannot be achieved without any minimum size condition (see, e.g., [28] and [20], Section 4, Remark 2.5). However, condition (43) might sometimes be too stringent for $d \ge 2$. For example, it rules out the case when $\mathbf{x}^* := (0.5, 0, \dots, 0)$ which means that the function class $\mathfrak{R}_1^d(c)$ excludes simple functions such as $f(\mathbf{x}) := \mathbb{I}\{x_1 \ge 1/2\}$. In Theorem A.1 (deferred to Appendix A of the Supplementary Material [21]), we show that when $d = 2$, it is possible to obtain the same risk bound under a weaker minimum size condition which does not rule out functions such as $f(\mathbf{x}) := \mathbb{I}\{x_1 \ge 1/2\}$.

The implication of Theorems 4.7 and A.1 is that there exists a subclass of $\mathfrak{R}^d$ consisting of indicators of upper right rectangles in $[0,1]^d$ over which the estimator $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$, when ideally tuned, achieves the near-parametric rate with some logarithmic factors. Simulations (see Section B.3 of the Supplementary Material [21]) indicate that this should also be true for a larger subclass of $\mathfrak{R}^d$ consisting of all functions in $\mathfrak{R}^d$ satisfying some minimum size condition, but our proof technique does not currently work in this generality. Ortelli and van de Geer [41] recently proved, for $d = 2$, near-parametric rates for the estimator (36) for a more general class of piecewise constant functions, but for a smaller loss function. Their proof technique is completely different from our approach.

Let us now briefly discuss the key ideas behind the proofs of Theorems 4.5, 4.6 and 4.7. Theorem 4.5 is proved via covering number arguments which relate $\mathcal{R}(\widehat{f}_{\mathrm{HK}\mathbf{0},V}, f^*)$ to covering numbers of $\{f : V_{\mathrm{HK}\mathbf{0}}(f) \le V\}$ and these covering numbers are controlled by invoking connections to distribution functions of signed measures. Theorem 4.6 is proved by Assouad's lemma with a multiscale construction of functions with bounded HK**0** variation. This multiscale construction is involved and taken from Blei et al. [7], Section 4.

The ideas for the proof of Theorem 4.7 (and also Theorem A.1) is borrowed from the proofs for the univariate case in Guntuboyina et al. [28] although the situation for $d \ge 2$ is much more complicated. At a high level, we use tangent cone connections where the goal is to control an appropriate size measure (Gaussian width) of the tangent cone of $\{f : V_{\mathrm{HK}\mathbf{0}}(f) \le V^*\}$ at f^*. This tangent cone can be explicitly computed (see Lemma C.11). To bound its Gaussian width, our key observation is that for functions f^* in $\mathfrak{R}_1^d(c)$, every element of the tangent cone can be broken down into lower-dimensional elements each of which is either nearly entirely monotone or has low HK**0** variation. The Gaussian width of the tangent cone can then be bounded by a combination of (suitably strengthened) versions of Theorem 4.4 and Theorem 4.5. This method unfortunately does not seem to work for arbitrary functions $f^* \in \mathfrak{R}^d$ because of certain technical issues which are mentioned in Remark C.2.

5. On the "dimension-independent" rate $n^{-2/3}$ in Theorem 4.1 and Theorem 4.5.

As mentioned previously, the dimension d appears in the bounds given by Theorem 4.1 and Theorem 4.5 only through the logarithmic term which means that $\widehat{f}_{\mathrm{EM}}$ and $\widehat{f}_{\mathrm{HK}\mathbf{0},V}$ attain "dimension-independent rates" ignoring logarithmic factors. We shall provide some insight and put these results in proper historical context in this section. In nonparametric statistics, it is well known that the rate of estimation of smooth functions based on n observations is $n^{-2m/(2m+d)}$ where d is the dimension and m is the order of smoothness [51]. The constraints of entire monotonicity and having finite HK**0** variation can be loosely viewed as smoothness constraints of order $m = d$. This is because, for smooth functions f, entire monotonicity is equivalent to

$$\frac{\partial^{|S|} f}{\prod_{j \in S} \partial x_j} \ge 0 \quad \text{for every } \varnothing \ne S \subseteq \{1, \dots, d\}$$

and the constraint of finite HK**0** variation is equivalent to

$$\frac{\partial^{|S|} f}{\prod_{j \in S} \partial x_j} \in L^1 \quad \text{for every } \varnothing \ne S \subseteq \{1, \dots, d\}. \tag{45}$$

Because derivatives of order d appear in these expressions, these constraints should be considered as smoothness constraints of order d. Note that taking $m = d$ in $n^{-2m/(2m+d)}$ gives $n^{-2/3}$.

Some other papers which studied such higher order constraints to obtain estimators having nearly dimension-free rates include [4, 13, 35, 40, 50, 59]. In particular, Lin [35] studied estimation under the constraint:

$$\frac{\partial^{|S|} f}{\prod_{j \in S} \partial x_j} \in L^2 \quad \text{for every } \varnothing \ne S \subseteq \{1, \dots, d\}. \tag{46}$$

The difference between (45) and (46) is that L^1 in (45) is replaced by L^2 in (46). Lin [35] proved that the minimax rate of convergence under (46) is $n^{-2/3}(\log n)^{2(d-1)/3}$ and constructed a linear estimator which is optimal over the class (46). Let us remark here that the L^2 constraint makes the class smaller compared to (45) and also enables linear estimators to achieve the optimal rate. However, linear estimators will not be optimal over $\{f : V_{\mathrm{HK}\mathbf{0}}(f) \le V\}$ as is well known in $d = 1$ (see Donoho and Johnstone [19]) and the estimator of Lin [35] will also not adapt to rectangular piecewise constant functions (note that it is not possible to extend (46) to nonsmooth functions in such a way that the constraint is satisfied by rectangular piecewise constant functions).

Let us also mention here that, in approximation theory, it is known that classes of smooth functions f on $[0, 1]^d$ satisfying mixed partial derivative constraints such as (45) or (46) allow one to overcome the curse of dimensionality to some extent from the perspective of metric entropy, approximation and interpolation (see, e.g., [9, 18, 52]).

Another way to impose higher order smoothness is to impose the constraint:

$$\frac{\partial^d f}{\partial x_j^d} \in L^1 \quad \text{for each } j = 1, \ldots, d \tag{47}$$

as in the Kronecker trend filtering method of order $k + 1 = d$ of Sadhanala et al. [50] who also proved that this leads to the dimension-free rate $n^{-2/3}$ up to logarithmic factors. There are some differences between the constraints (45) and (47). For example, product functions $f(x_1, \ldots, x_d) := f_1(x_1) \ldots f_d(x_d)$ satisfy (45) provided each f_j satisfies $f_j' \in L_1$ while they will satisfy (47) provided $f_j^{(d)} \in L_1$.

Finally, let us mention that, in the usual multivariate extensions of isotonic regression and total variation denoising, one uses partial derivatives only of the first order which leads to rates of convergence that are exponential in the dimension d. For example, the usual multivariate isotonic regression (see, e.g., Robertson et al. [46], Section 1.3) considers the class $\mathcal{F}_{\mathrm{M}}^d$ of multivariate monotone functions which only imposes first order constraints. The rate of convergence here is given by $n^{-1/d}$ as recently shown in Han et al. [30]. This rate is exponentially slow in the dimension d. One sees the same rate behavior for the multivariate total variation denoising estimator (which also imposes only first-order constraints) originally proposed by Rudin et al. [47] and whose theoretical behavior is studied in Chatterjee and Goswami [11], Hütter and Rigollet [32], Ortelli and van de Geer [43], Ruiz et al. [48] and Sadhanala et al. [49].

Acknowledgments. We are extremely thankful to the Associate Editor and the two anonymous referees for several insightful comments that led to many improvements in the paper. We are also thankful to Frank Fuchang Gao for clarifying some technical arguments in the paper [7], to Ming Yuan for informing us of the paper [35] and to Jake Soloff for helpful comments.

The first author was supported by a NSF Graduate Research Fellowship. The second author was supported by NSF CAREER Grant DMS-1654589. The third author was supported by NSF Grants DMS-1712822 and AST-1614743.

SUPPLEMENTARY MATERIAL

Supplement to "Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation" (DOI: 10.1214/20-AOS1977SUPP; .pdf). This contains additional results and simulations as well as all the proofs of the results of the paper.

REFERENCES

[1] 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

[2] AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Stat.* **26** 641–647. MR0073895 https://doi.org/10.1214/aoms/1177728423

[3] BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression*. Wiley, London-Sydney. Wiley Series in Probability and Mathematical Statistics. MR0326887

[4] BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* **39** 930–945. MR1237720 https://doi.org/10.1109/18.256500

[5] BELLEC, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383 https://doi.org/10.1214/17-AOS1566

[6] BENKESER, D. and VAN DER LAAN, M. (2016). The highly adaptive lasso estimator. In 2016 *IEEE International Conference on Data Science and Advanced Analytics* (*DSAA*) 689–696. IEEE.

[7] 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

[8] BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Stat.* **26** 607–616. MR0073894 https://doi.org/10.1214/aoms/1177728420

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

[10] CHAMBOLLE, A., CASELLES, V., CREMERS, D., NOVAGA, M. and POCK, T. (2010). An introduction to total variation for image analysis. In *Theoretical Foundations and Numerical Methods for Sparse Recovery. Radon Ser. Comput. Appl. Math.* **9** 263–340. de Gruyter, Berlin. MR2731599 https://doi.org/10.1515/9783110226157.263

[11] CHATTERJEE, S. and GOSWAMI, S. (2019). New risk bounds for 2d total variation denoising. arXiv preprint. Available at arXiv:1902.01215.

[12] CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878 https://doi.org/10.1214/15-AOS1324

[13] CHKIFA, A., DEXTER, N., TRAN, H. and WEBSTER, C. G. (2018). Polynomial approximation via compressed sensing of high-dimensional functions on lower sets. *Math. Comp.* **87** 1415–1450. MR3766392 https://doi.org/10.1090/mcom/3272

[14] CONDAT, L. (2013). A direct algorithm for 1-d total variation denoising. *IEEE Signal Process. Lett.* **20** 1054–1057.

[15] CONDAT, L. (2017). Discrete total variation: New definition and minimization. *SIAM J. Imaging Sci.* **10** 1258–1290. MR3684410 https://doi.org/10.1137/16M1075247

[16] DALALYAN, A. S., HEBIRI, M. and LEDERER, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23** 552–581. MR3556784 https://doi.org/10.3150/15-BEJ756

[17] DENG, H. and ZHANG, C.-H. (2018). Isotonic regression in multi-dimensional spaces and graphs. arXiv preprint. Available at arXiv:1812.08944.

[18] DONOHO, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math. Challenges Lect.* **1** 375.

[19] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414 https://doi.org/10.1214/aos/1024691081

[20] FAN, Z. and GUAN, L. (2018). Approximate ℓ_0-penalized estimation of piecewise-constant signals on graphs. *Ann. Statist.* **46** 3217–3245. MR3852650 https://doi.org/10.1214/17-AOS1656

[21] FANG, B., GUNTUBOYINA, A. and SEN, B. (2021). Supplement to "Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation." https://doi.org/10.1214/20-AOS1977SUPP

[22] FELLER, W. (2015). Completely monotone functions and sequences. In *Selected Papers I* 497–510. Springer, Berlin.

[23] GAO, F. (2013). Bracketing entropy of high dimensional distributions. In *High Dimensional Probability VI. Progress in Probability* **66** 3–17. Birkhäuser/Springer, Basel. MR3443489

[24] GAO, F., LI, W. V. and WELLNER, J. A. (2010). How many Laplace transforms of probability measures are there? *Proc. Amer. Math. Soc.* **138** 4331–4344. MR2680059 https://doi.org/10.1090/S0002-9939-2010-10448-3

[25] GILL, R. D., VAN DER LAAN, M. J. and WELLNER, J. A. (1995). Inefficient estimators of the bivariate survival function for three models. *Ann. Inst. Henri Poincaré Probab. Stat.* **31** 545–597. MR1338452

[26] GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints*. *Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. Estimators, algorithms and asymptotics. MR3445293 https://doi.org/10.1017/CBO9781139020893

[27] GUNTUBOYINA, A. and SEN, B. (2018). Nonparametric shape-restricted regression. *Statist*. *Sci*. **33** 568–594. MR3881209 https://doi.org/10.1214/18-STS665

[28] 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

[29] GUO, D. and WANG, X. (2006). Quasi-Monte Carlo filtering in nonlinear dynamic systems. *IEEE Trans*. *Signal Process*. **54** 2087–2098.

[30] HAN, Q., WANG, T., CHATTERJEE, S. and SAMWORTH, R. J. (2019). Isotonic regression in general dimensions. *Ann*. *Statist*. **47** 2440–2471. MR3988762 https://doi.org/10.1214/18-AOS1753

[31] HOBSON, E. W. (1958). *The Theory of Functions of a Real Variable and the Theory of Fourier's Series*. *Vol*. *I*. Dover, New York, NY. MR0092828

[32] HÜTTER, J.-C. and RIGOLLET, P. (2016). Optimal rates for total variation denoising. In *Conference on Learning Theory* 1115–1146.

[33] 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

[34] LEONOV, A. S. (1996). On the total variation for functions of several variables and a multidimensional analog of Helly's selection principle. *Math*. *Notes* **63** 61–71.

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

[36] LIN, K., SHARPNACK, J. L., RINALDO, A. and TIBSHIRANI, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems* 6884–6893.

[37] 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

[38] MEYER, M. and WOODROOFE, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann*. *Statist*. **28** 1083–1104. MR1810920 https://doi.org/10.1214/aos/1015956708

[39] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math*. **1738** 85–277. Springer, Berlin. MR1775640

[40] NIYOGI, P. and GIROSI, F. (1999). Generalization bounds for function approximation from scattered noisy data. *Adv*. *Comput*. *Math*. **10** 51–80. MR1671871 https://doi.org/10.1023/A:1018966213079

[41] ORTELLI, F. and VAN DE GEER, S. (2018). On the total variation regularized estimator over a class of tree graphs. *Electron*. *J*. *Stat*. **12** 4517–4570. MR3892703 https://doi.org/10.1214/18-ejs1519

[42] ORTELLI, F. and VAN DE GEER, S. (2019). Oracle inequalities for image denoising with total variation regularization. arXiv preprint. Available at arXiv:1911.07231.

[43] ORTELLI, F. and VAN DE GEER, S. (2019). Synthesis and analysis in total variation regularization. arXiv preprint. Available at arXiv:1901.06418.

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

[45] PRAUSE, A. and STELAND, A. (2017). Sequential detection of three-dimensional signals under dependent noise. *Sequential Anal*. **36** 151–178. MR3665833 https://doi.org/10.1080/07474946.2017.1319674

[46] ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. MR0961262

[47] RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. Experimental mathematics: Computational issues in nonlinear science (Los Alamos, NM, 1991). *Phys*. *D* **60** 259–268. MR3363401 https://doi.org/10.1016/0167-2789(92)90242-F

[48] RUIZ, D Á, M., LI, H. and MUNK, A. (2018). Frame-constrained total variation regularization for white noise regression. arXiv preprint. Available at arXiv:1807.02038.

[49] SADHANALA, V., WANG, Y.-X. and TIBSHIRANI, R. J. (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems* 3513–3521.

[50] SADHANALA, V., WANG, Y.-X., SHARPNACK, J. L. and TIBSHIRANI, R. J. (2017). Higher-order total variation classes on grids: Minimax theory and trend filtering methods. In *Advances in Neural Information Processing Systems* 5800–5810.

[51] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann*. *Statist*. **10** 1040–1053. MR0673642

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

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

[54] VAN DER LAAN, M. (2017). Finite sample inference for targeted learning. arXiv preprint. Available at arXiv:1708.09502.

[55] VAN DER LAAN, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive Lasso. *Int*. *J*. *Biostat*. **13** 20150097. MR3724476 https://doi.org/10.1515/ijb-2015-0097

[56] VAN DER LAAN, M. J., BENKESER, D. and CAI, W. (2019). Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. arXiv preprint. Available at arXiv:1908.05607.

[57] VAN DER LAAN, M. J. and BIBAUT, A. F. (2017). Uniform consistency of the highly adaptive lasso estimator of infinite dimensional parameters. arXiv preprint. Available at arXiv:1709.06256.

[58] VAPNIK, V. N. and CHERVONENKIS, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of Complexity* 11–30. Springer, Cham. Reprint of Theor. Probability Appl. **1**(6) (1971), 264–280. MR3408730

[59] WAHBA, G., WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann*. *Statist*. **23** 1865–1895. MR1389856 https://doi.org/10.1214/aos/1034713638

[60] WIDDER, D. V. (1941). *The Laplace Transform*. *Princeton Mathematical Series*, *V.* 6. Princeton Univ. Press, Princeton, NJ. MR0005923

[61] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann*. *Statist*. **27** 1564–1599. MR1742500 https://doi.org/10.1214/aos/1017939142

[62] YOUNG, W. H. and YOUNG, G. C. (1924). On the Discontinuties of Monotone Functions of Several Variables. *Proc*. *Lond*. *Math*. *Soc*. (2) **22** 124–142. MR1575698 https://doi.org/10.1112/plms/s2-22.1.124

[63] ZHANG, C.-H. (2002). Risk bounds in isotonic regression. *Ann*. *Statist*. **30** 528–555. MR1902898 https://doi.org/10.1214/aos/1021379864

[64] ZHANG, T. (2019). Element-wise estimation error of a total variation regularized estimator for change point detection. arXiv preprint. Available at arXiv:1901.00914.

[65] ZIEMER, W. P. (2012). *Weakly Differentiable Functions*: *Sobolev Spaces and Functions of Bounded Variation* **120**. Springer, Berlin.