The Annals of Statistics
2020, Vol. 48, No. 6, 3672–3698
https://doi.org/10.1214/20-AOS1947
© Institute of Mathematical Statistics, 2020

ISOTONIC REGRESSION IN MULTI-DIMENSIONAL SPACES AND GRAPHS

BY HANG DENG[*] AND CUN-HUI ZHANG[†]

Department of Statistics, Rutgers University, [*]*hdeng@stat.rutgers.edu;* [†]*czhang@stat.rutgers.edu*

In this paper, we study minimax and adaptation rates in general isotonic regression. For uniform deterministic and random designs in $[0,1]^d$ with $d \geq 2$ and $N(0,1)$ noise, the minimax rate for the ℓ_2 risk is known to be bounded from below by $n^{-1/d}$ when the unknown mean function f is nondecreasing and its range is bounded by a constant, while the least squares estimator (LSE) is known to nearly achieve the minimax rate up to a factor $(\log n)^\gamma$ where n is the sample size, $\gamma = 4$ in the lattice design and $\gamma = \max\{9/2, (d^2+d+1)/2\}$ in the random design. Moreover, the LSE is known to achieve the adaptation rate $(K/n)^{-2/d}\{1 \vee \log(n/K)\}^{2\gamma}$ when f is piecewise constant on K hyperrectangles in a partition of $[0,1]^d$.

Due to the minimax theorem, the LSE is identical on every design point to both the max-min and min-max estimators over all upper and lower sets containing the design point. This motivates our consideration of estimators which lie in-between the max-min and min-max estimators over possibly smaller classes of upper and lower sets, including a subclass of block estimators. Under a qth moment condition on the noise, we develop ℓ_q risk bounds for such general estimators for isotonic regression on graphs. For uniform deterministic and random designs in $[0,1]^d$ with $d \geq 3$, our ℓ_2 risk bound for the block estimator matches the minimax rate $n^{-1/d}$ when the range of f is bounded and achieves the near parametric adaptation rate $(K/n)\{1 \vee \log(n/K)\}^d$ when f is K-piecewise constant. Furthermore, the block estimator possesses the following oracle property in variable selection: When f depends on only a subset S of variables, the ℓ_2 risk of the block estimator automatically achieves up to a poly-logarithmic factor the minimax rate based on the oracular knowledge of S.

1. Introduction.

Let $G = (V, E)$ be a directed graph with vertex set V and edge set E. For $\boldsymbol{a}$ and $\boldsymbol{b}$ in V, we say that $\boldsymbol{a}$ is a descendant of $\boldsymbol{b}$ if E contains a chain of edges from $\boldsymbol{v}_j$ to $\boldsymbol{v}_{j+1}$ such that $\boldsymbol{b} = \boldsymbol{v}_0$ and $\boldsymbol{a} = \boldsymbol{v}_m$ for some finite $m \geq 0$. We write $\boldsymbol{a} \preceq \boldsymbol{b}$ if $\boldsymbol{a} = \boldsymbol{b}$ or $\boldsymbol{a}$ is a descendant of $\boldsymbol{b}$. A function $f : V \to \mathbb{R}$ is nondecreasing on the graph G if $f(\boldsymbol{a}) \leq f(\boldsymbol{b})$ whenever $\boldsymbol{a} \preceq \boldsymbol{b}$. Let $\mathcal{F}$ be the class of all nondecreasing functions on G. In isotonic regression, we observe $\boldsymbol{x}_i \in V$ and $y_i \in \mathbb{R}$ satisfying

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \text{for some } f \in \mathcal{F}, \tag{1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent noise variables with $\mathbb{E}\varepsilon_i = 0$ and $\mathrm{Var}(\varepsilon_i) \leq \sigma^2$ given the (deterministic or random) design points $\{\boldsymbol{x}_i\}$. Note that we allow $|V| > n$.

An interesting special case of (1) is the multiple isotonic regression where $V \subset \mathbb{R}^d$ is a subset of a certain Euclidean space of dimension d, and for $\boldsymbol{a} = (a_1, \ldots, a_d)^T \in \mathbb{R}^d$ and $\boldsymbol{b} = (b_1, \ldots, b_d)^T \in \mathbb{R}^d$, $\boldsymbol{a} \preceq \boldsymbol{b}$ iff $a_j \leq b_j$ for all $1 \leq j \leq d$. In this case, $\mathcal{F}$ is the class of all nondecreasing functions on V.

Received December 2018; revised October 2019.

MSC2020 subject classifications. 62G05, 62G08, 62C20.

Key words and phrases. Isotonic regression, multiple isotonic regression, isotonic regression on graphs, max-min estimator, min-max estimator, block estimator, lattice design, random design, minimax rate, adaptive estimation, variable selection, oracle property.

Let $\boldsymbol{f}_n = (f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n))^T$ and $\widehat{\boldsymbol{f}}_n = (\widehat{f}_n(\boldsymbol{x}_1), \dots, \widehat{f}_n(\boldsymbol{x}_n))^T$ for any estimator $\widehat{f}_n$ of f. We are interested in the estimation of f under the (normalized) ℓ_q risk

$$R_q(\widehat{\boldsymbol{f}}_n, \boldsymbol{f}_n) = \frac{1}{n}\mathbb{E}\|\widehat{\boldsymbol{f}}_n - \boldsymbol{f}_n\|_q^q = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\widehat{f}_n(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)|^q. \tag{2}$$

In this case, a specification of $\widehat{\boldsymbol{f}}_n$ is sufficient for the definition of $\widehat{f}_n$. For multiple isotonic regression with random design in $V \subseteq \mathbb{R}^d$, we are also interested in the L_q risk

$$R_q^*(\widehat{f}_n, f) = \mathbb{E}\|\widehat{f}_n - f\|_{L_q(V)}^q = \mathbb{E}\int_V |\widehat{f}_n(\boldsymbol{x}) - f(\boldsymbol{x})|^q \, d\boldsymbol{x}. \tag{3}$$

The literature of univariate isotonic regression ($d = 1$) encompasses at least the past six decades; see, for example, Brunk (1955), Ayer et al. (1955), Grenander (1956), Prakasa Rao (1969), Groeneboom (1985), van de Geer (1990, 1993), Donoho (1990), Birgé and Massart (1993), Woodroofe and Sun (1993), Wang (1996), Durot (2007, 2008) and Yang and Barber (2019) among many others for some key developments. The least squares estimator (LSE), say $\widehat{f}_n^{(\mathrm{lse})}$, has been the focus of this literature. We describe in some detail here existing results on minimax and adaptation rates as they are directly related to our study. For any $a < b$, the ℓ_q risk of the LSE in the interval $[a, b]$ is bounded by

$$\mathbb{E}\sum_{a \le x_i \le b} |\widehat{f}_n^{(\mathrm{lse})}(x_i) - f(x_i)|^q \le C_q\sigma^q\left\{n_{a,b}\left(\frac{\Delta_{a,b}(\boldsymbol{f}_n/\sigma)}{n_{a,b}} \wedge 1\right)^{q/3} + \sum_{j=1}^{n_{a,b}} j^{-q/2}\right\}, \tag{4}$$

where $\Delta_{a,b}(\boldsymbol{f}_n/\sigma) = \max_{a \le x_i < x_j \le b}\{f(x_j) - f(x_i)\}/\sigma$ is the range-to-noise ratio for the mean vector $\boldsymbol{f}_n$ in $[a, b]$, $n_{a,b} = \#\{j : a \le x_j \le b\}$ is the number of design points in the interval, and C_q is a constant depending on q only. This result can be found in Meyer and Woodroofe (2000) for $n_{a,b} = n$, $q = 2$ and $\varepsilon_i \sim N(0, \sigma^2)$, and in Zhang (2002) for general $a < b$ and $1 \le q < 3$ under a $(q \vee 2)$th moment condition on ε_i. For $\Delta_{-\infty,\infty}(\boldsymbol{f}_n/\sigma) \le \Delta_n^* \asymp 1$, (4) yields the cube-root rate $\sigma^q(\Delta_n^*/n)^{q/3}$ for the LSE in terms of the ℓ_q risk in (2). By summing over the risk bound (4) over K intervals $[a_k, b_k]$ with $\Delta_{a_k,b_k}(\boldsymbol{f}_n/\sigma) = 0$, the LSE can be seen to achieve the near parametric adaptation rate $(K/n)\{1 \vee \log(n/K)\}$ in the mean squared risk when the unknown f is piecewise constant on the K intervals and $x_i \in \bigcup_{k=1}^{K}[a_k, b_k]$ for all $i \le n$. This adaptation rate was explicitly given in Chatterjee, Guntuboyina and Sen (2015). However, Gao, Han and Zhang (2017) proved that the sharp adaptation rate in the mean squared risk, achieved by a penalized LSE, is $(K/n)\log\log(16n/K)$ in the piecewise constant case. Moreover, by summing over the risk bound (4) over a growing number of disjoint intervals, the LSE has been shown to converge faster than the cube root rate when the measure $f(dx)$ is singular to the Lebesgue measure (Zhang (2002)).

Compared with the rich literature on univariate isotonic regression, our understanding of the multiple isotonic regression, that is, $V \subset \mathbb{R}^d$ with $d > 1$, is quite limited. A major difficulty is that the design points are typically only partially ordered. Univariate risk bounds can be directly applied to linearly ordered paths in V, but this typically does not yield a nearly minimax rate. However, significant advances have been made recently on the minimax and adaptation rates for the LSE. For $n_1 \times \cdots \times n_d$ lattice designs with $n = \prod_{j=1}^d n_j$, the LSE provides

$$R_2(\widehat{\boldsymbol{f}}_n^{(\mathrm{lse})}, \boldsymbol{f}_n) \le C_d\sigma^2\{\Delta(\boldsymbol{f}_n/\sigma)n^{-1/d}(\log n)^{\gamma} + n^{-2/d}(\log n)^{2\gamma}\} \tag{5}$$

in certain settings, where $\Delta(\boldsymbol{f}_n/\sigma) = \max_{1 \le i < j \le n}|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)|/\sigma$ is the range-to-noise ratio of the mean over the design points. For Gaussian ε_i and $n_1 = \cdots = n_d$, the minimax rate is bounded from below by

$$\inf_{\widehat{\boldsymbol{f}}_n}\sup_{\Delta(\boldsymbol{f}_n/\sigma) \le \Delta_n^*} R_2(\widehat{\boldsymbol{f}}_n, \boldsymbol{f}_n) \ge \sigma^2\min\{1, C_0 n^{-1/d}\Delta_n^*\}. \tag{6}$$

Moreover, when f is piecewise constant on K hyperrectangles in a partition of the lattice,

$$R_2(\widehat{\boldsymbol{f}}_n^{(\mathrm{lse})}, \boldsymbol{f}_n) \le C_d \sigma^2 (K/n)^{2/d} \{1 \vee \log(n/K)\}^{2\gamma}. \tag{7}$$

For $d = 2$ and Gaussian noise, Chatterjee, Guntuboyina and Sen (2018) proved the above mean squared risk bounds with $\gamma = 4$. Thus, up to a logarithmic factor, the LSE is nearly rate minimax for a wide range of Δ_n^* and also nearly adaptive to the parametric rate $\sigma^2 K/n$ when f is piecewise constant on K rectangles. Han et al. (2019) extended the results of Chatterjee, Guntuboyina and Sen (2018) from $d = 2$ to $d > 2$ under the conditions $n_1 = \cdots = n_d$ and $\Delta(\boldsymbol{f}_n/\sigma) \le \Delta_n^* = 1$ in (5) and (6), and also proved parallel results for random designs with a larger $\gamma = \max\{9/2, (d^2+d+1)/2\}$. However, there is still a gap of a poly-logarithmic factor between such upper and lower minimax bounds for $d \ge 2$, and it is still unclear from (7) the feasibility of near adaptation to the parametric rate $\sigma^2 K/n$ for $d \ge 3$ when f is piecewise constant on K hyperrectangles.

We have also seen some progresses in adaptive estimation to variable selection in isotonic regression on lattices with $\max_{j \le d} n_j \le C_d n^{1/d}$. When the unknown mean function depends on only a *known* subset of s variables, say $f(\boldsymbol{x}) = f_S(\boldsymbol{x}_S)$ where $\boldsymbol{x}_S = (x_j, j \in S)^T$ with $|S| = s$, one may use the LSE, say $\widehat{f}_{n,S}^{(\mathrm{lse})}$, based on the average of y_i given $\boldsymbol{x}_S$ to attain

$$R_2(\widehat{\boldsymbol{f}}_{n,S}^{(\mathrm{lse})}, \boldsymbol{f}_n) \le \begin{cases} C_d \sigma_S^2 [\Delta(\boldsymbol{f}_n/\sigma_S) n^{-1/d} (\log n)^{\gamma} + n^{-2/d} (\log n)^{2\gamma}], & s \ge 2, \\ C_d \sigma_S^2 [\{(\Delta(\boldsymbol{f}_n/\sigma_S) n^{-1/d}) \wedge 1\}^{2/3} + n^{-1/d} \log n], & s = 1, \end{cases} \tag{8}$$

with $\sigma_S^2 = \sigma^2 / \prod_{j \notin S} n_j \le C_d \sigma^2 / n^{1-s/d}$, which would match the minimax rate for Gaussian ε_i for a proper range of $\Delta(\boldsymbol{f}_n/\sigma_S)$ as we discussed in the previous paragraph. For unknown S with $d \ge 2$ and $\Delta(\boldsymbol{f}_n/\sigma) \le 1 = \sigma$, Han et al. (2019) proved that the LSE $\widehat{f}_n^{(\mathrm{lse})}$ for the general f automatically achieves the rate $n^{-4/(3d)}(\log n)^{16/3}$ for $s = d - 1$ and $n^{-2/d}(\log n)^8$ for $s \le d - 2$. As $\Delta(\boldsymbol{f}_n/\sigma_S) \asymp n^{(d-s)/(2d)}$ in their setting, (8) would yield the rates $n^{-(d-s)/(2d)-1/d}$ for $s \ge 2$ and $n^{-(d-1)/d-(3-d)_+/(3d)}$ for $s = 1$ up to a logarithmic factor. These oracle minimax rates nearly match the adaptation rates in Han et al. (2019) for $d - s = 2$ or $(d, s) = (2, 1)$, but not for other configurations of (d, s).

We consider isotonic regression on directed graphs, that is, with general domain V in (1), including $V \subset \mathbb{R}^d$ as a special case. In this general setting, Robertson, Wright and Dykstra (1988) proved the following minimax formula for the LSE on the design points:

$$\widehat{f}_n^{(\mathrm{lse})}(\boldsymbol{x}) = \max_{U \ni \boldsymbol{x}} \min_{L \ni \boldsymbol{x}} \overline{y}_{U \cap L} = \min_{L \ni \boldsymbol{x}} \max_{U \ni \boldsymbol{x}} \overline{y}_{U \cap L} \tag{9}$$

for $\boldsymbol{x} = \boldsymbol{x}_i$, $i = 1, \ldots, n$, where the maximum is taken over all upper sets U containing $\boldsymbol{x}$, the minimum over all lower sets L containing $\boldsymbol{x}$, and $\overline{y}_A$ is the average of the observed y_i over $\boldsymbol{x}_i \in A$ for any $A \subseteq V$. As the high complexity of the upper and lower sets for $d \ge 2$ could be the culprit behind the possible suboptimal performance of the LSE in convergence and adaptation rates, we consider a class of block estimators involving rectangular upper and lower sets. As the minimax theorem no longer holds in this setting in general, the block estimator, say $\widehat{f}_n^{(\mathrm{block})}(\boldsymbol{x})$, is defined as any estimator in-between the following max-min and min-max estimators:

$$\begin{aligned} \widehat{f}_n^{(\text{max-min})}(\boldsymbol{x}) &= \max_{\boldsymbol{u} \preceq \boldsymbol{x}, n_{\boldsymbol{u},*} > 0} \min_{\boldsymbol{x} \preceq \boldsymbol{v}, n_{\boldsymbol{u},\boldsymbol{v}} > 0} \overline{y}_{[\boldsymbol{u},\boldsymbol{v}]} \quad \forall \boldsymbol{x} \in V, \\ \widehat{f}_n^{(\text{min-max})}(\boldsymbol{x}) &= \min_{\boldsymbol{x} \preceq \boldsymbol{v}, n_{*,\boldsymbol{v}} > 0} \max_{\boldsymbol{u} \preceq \boldsymbol{x}, n_{\boldsymbol{u},\boldsymbol{v}} > 0} \overline{y}_{[\boldsymbol{u},\boldsymbol{v}]} \quad \forall \boldsymbol{x} \in V, \end{aligned} \tag{10}$$

where $[\boldsymbol{u}, \boldsymbol{v}] = \{\boldsymbol{x} : \boldsymbol{u} \preceq \boldsymbol{x} \preceq \boldsymbol{v}\}$, $n_{\boldsymbol{u},\boldsymbol{v}} = \#\{i \le n : \boldsymbol{x}_i \in [\boldsymbol{u}, \boldsymbol{v}]\}$, $n_{\boldsymbol{u},*} = \#\{i \le n : \boldsymbol{u} \preceq \boldsymbol{x}_i\}$ and $n_{*,\boldsymbol{v}} = \#\{i \le n : \boldsymbol{x}_i \preceq \boldsymbol{v}\}$. The idea of replacing the general level sets $U \cap L$ by rectangular blocks $[\boldsymbol{u}, \boldsymbol{v}]$ is not new as a preliminary version of the block estimator in the case of $V =$

$[0,1]^d$ was considered in Fokianos, Leucht and Neumann (2017). Some more delicate details of different versions of the block estimator are discussed in Section 2.

We derive in Section 3 a general ℓ_q risk bound for the above block estimator on graphs. For $n_1 \times \cdots \times n_d$ lattice designs with $d \geq 2$, our general risk bound yields

$$R_2(\widehat{\boldsymbol{f}}_n^{(\text{block})}, \boldsymbol{f}_n) \leq C_d \sigma^2 \min\{1, \Delta(\boldsymbol{f}_n/\sigma) n^{-1/d} (\log n)^{I\{d=2\}} + n^{-1} (\log n)^d\} \tag{11}$$

when $\max_{j \leq d} n_j \leq C_d n^{1/d}$, compared with (5) and (6), and the adaptation rate

$$R_2(\widehat{\boldsymbol{f}}_n^{(\text{block})}, \boldsymbol{f}_n) \leq C_d \sigma^2 (K/n)\{1 \vee \log(n/K)\}^d \tag{12}$$

when the true f is nondecreasing and piecewise constant on K hyperrectangles, compared with (7).

We also explore the phase transition of the risk bounds, both the minimax lower bound and the upper risk bound for the block estimator, by presenting them using its effective dimension s in the sense that the risk bound only depends on the largest s n_j's. For example, when $n_1 \geq n_2 \geq \cdots \geq n_d$ and $n_2^{3/2}/n_1^{1/2} \leq \Delta(\boldsymbol{f}_n/\sigma)$, we show that the risk bound for the block estimator in d-dimensional isotonic regression with n design points is almost no different from that in univariate isotonic regression with n_1 design points. This phase transition, captured by effective dimension, proved for $d = 2$ in Chatterjee, Guntuboyina and Sen (2018), is new for $d > 2$.

Moreover, perhaps more interestingly, we prove that when the unknown f depends on an unknown set of s variables, the block estimator achieves near adaptation to the oracle selection in the sense that for $\Delta(\boldsymbol{f}_n/\sigma) \leq \Delta_n^*$,

$$\begin{aligned} & R_2(\widehat{\boldsymbol{f}}_n^{(\text{block})}, \boldsymbol{f}_n) \\ & \leq \begin{cases} C_d \sigma_S^2 \min\big[(\log n)^{d-s}, \Delta_n^* n^{\frac{d-s-2}{2d}} (\log n)^{I\{s=2\}} + n^{-\frac{s}{d}} (\log n)^d\big], & s \geq 2, \\ C_d \sigma_S^2 \min\big[(\log n)^{d-1}, (\Delta_n^* n^{\frac{d-s-2}{2d}})^{2/3} + n^{-\frac{1}{d}} (\log n)^d\big], & s = 1, \end{cases} \end{aligned} \tag{13}$$

with $\sigma_S^2 = \sigma^2 / \prod_{j \notin S} n_j \leq C_d \sigma^2 / n^{1-s/d}$, while the oracle minimax rate with the knowledge of S is bounded from below by

$$\begin{aligned} & \inf_{\widehat{\boldsymbol{f}}_n} \sup_{\boldsymbol{f}_n} \{R_2(\widehat{\boldsymbol{f}}_n, \boldsymbol{f}_n) : \boldsymbol{f}_n \in \mathcal{F}_n, f(\boldsymbol{x}) = f_S(\boldsymbol{x}_S), \Delta(\boldsymbol{f}_n/\sigma) \leq \Delta_n^*\} \\ & \geq \begin{cases} C_d \sigma^2 n^{-1+s/d} \min\big[1, \Delta_n^* n^{(d-s-2)/(2d)}\big], & s \geq 2, \\ C_d \sigma^2 n^{-1+1/d} \min\big[1, (\Delta_n^* n^{(d-3)/(2d)})^{2/3}\big], & s = 1, \end{cases} \end{aligned} \tag{14}$$

where $\mathcal{F}_n = \{\boldsymbol{f}_n : f \in \mathcal{F}\}$.

Let $\overline{\boldsymbol{f}}_n^*$ be the noiseless version of the block estimator. When the isotonic regression model is misspecified in the sense of having a nonmonotone regression function, we prove that the error bounds discussed above still hold if $\overline{\boldsymbol{f}}_n^*$ is treated as the estimation target; (11), (12) and (13) are valid with $\boldsymbol{f}_n$ replaced by $\overline{\boldsymbol{f}}_n^*$ when $f \notin \mathcal{F}$ in (1). However, such results are of a less ideal form compared with the existing oracle inequalities for the LSE under misspecified monotonicity assumption (Chatterjee, Guntuboyina and Sen (2015, 2018), Bellec (2018), Han et al. (2019)).

We summarize our main results as follows. In terms of the mean squared risk, the block estimator is rate minimax for $\Delta(\boldsymbol{f}_n/\sigma) \leq \Delta_n^*$ with a wide range of Δ_n^* (with no extra logarithmic factor for $d \neq 2$), achieves near parametric adaptation in the piecewise constant case, and also achieves near adaptation to the oracle minimax rate in variable selection. Furthermore, we prove parallel results for the integrated risk for i.i.d. random designs in $[0,1]^d$ when the joint density of the design point is uniformly bounded away from zero and infinity. In addition to Sections 2 and 3, we present in Section 4 some simulation results to demonstrate the advantage of the block estimator over the LSE in multiple isotonic regression. The full proofs of all theorems, propositions and lemmas in this paper are relegated to the Supplementary Material (Deng and Zhang (2020)).

Here and in the sequel, the following notation is used. For $\{\boldsymbol{a}, \boldsymbol{b}\} \subset V$, we say $\boldsymbol{b}$ is larger than $\boldsymbol{a}$ when $\boldsymbol{a} \preceq \boldsymbol{b}$, and we set $[\boldsymbol{a}, \boldsymbol{b}] = \{\boldsymbol{x} \in V : \boldsymbol{a} \preceq \boldsymbol{x} \preceq \boldsymbol{b}\}$ as a block in $G = (V, E)$. We denote by n_A the number of sampled points in A, that is, $n_A = \#\{i \le n : \boldsymbol{x}_i \in A\}$, and set $n_{\boldsymbol{a},\boldsymbol{b}} = n_{[\boldsymbol{a},\boldsymbol{b}]}$, $n_{\boldsymbol{a},*} = \#\{i \le n : \boldsymbol{a} \preceq \boldsymbol{x}_i\}$ and $n_{*,\boldsymbol{b}} = \#\{i \le n : \boldsymbol{x}_i \preceq \boldsymbol{b}\}$. For $\boldsymbol{a} = (a_1, \dots, a_d)^T \in \mathbb{R}^d$ and $\boldsymbol{b} = (b_1, \dots, b_d)^T \in \mathbb{R}^d$, $\boldsymbol{a} \preceq \boldsymbol{b}$ iff $a_j \le b_j$ for all $1 \le j \le d$, and this is also expressed as $\boldsymbol{a} \le \boldsymbol{b}$. We denote by C a positive numerical constant, and C_{index} a positive constant depending on the "index" only. For example, $C_{q,d}$ is a positive constant depending on (q, d) only. For the sake of convenience, the value of such a constant with the same subscript may change from one appearance to the next. We may write $x \lesssim_{\text{index}} y$ when $x \le C_{\text{index}} y$. Finally, we set $\log_+(x) = 1 \vee \log x$.

2. The least squares and block estimators.

Given design points $\boldsymbol{x}_i \in V$ and responses $y_i \in \mathbb{R}$, the isotonic LSE is formally defined as

$$\widehat{f}_n^{(\text{lse})} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^n \{y_i - f(\boldsymbol{x}_i)\}^2,$$

where $\mathcal{F} = \{f : f(\boldsymbol{u}) \le f(\boldsymbol{v})\ \forall \boldsymbol{u} \preceq \boldsymbol{v}\}$ is the set of all nondecreasing functions on the directed graph $G = (V, E)$. As the squared loss only involves the value of f at the design points, this LSE is any nondecreasing extension of the LSE of the mean vector $\boldsymbol{f}_n = (f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n))^T$ in (1),

$$\widehat{\boldsymbol{f}}_n^{(\text{lse})} = \arg\min_{\boldsymbol{f}_n \in \mathcal{F}_n} \|\boldsymbol{y} - \boldsymbol{f}_n\|_2^2, \tag{15}$$

where $\boldsymbol{y} = (y_1, \dots, y_n)^T$ and $\mathcal{F}_n = \{\boldsymbol{f}_n : f \in \mathcal{F}\} \subset \mathbb{R}^n$. As $\mathcal{F}_n$ is defined with no more than $\binom{n}{2}$ linear constraints, $\widehat{\boldsymbol{f}}_n^{(\text{lse})}$ can be computed with quadratic programming. Potentially more efficient algorithms for the LSE have been developed in Dykstra (1983), Kyng, Rao and Sachdeva (2015) and Stout (2015), among others.

As mentioned in the Introduction, the LSE $\widehat{\boldsymbol{f}}_n^{(\text{lse})}$ has an explicit representation in the minimax formula (9) for isotonic regression on graphs in general (Robertson, Wright and Dykstra (1988)), although this fact is better known in the univariate case. As the high complexity of the general upper and lower sets in the minimax formula seems to be the cause of the analytical or possibly real gap between the risk of the LSE and the optimal minimax and adaptation rates, we consider in this paper block estimators $\widehat{f}_n^{(\text{block})}$ of the form

$$\begin{aligned} &\min\{\widehat{f}_n^{(\text{max-min})}(\boldsymbol{x}), \widehat{f}_n^{(\text{min-max})}(\boldsymbol{x})\} \\ &\quad \le \widehat{f}_n^{(\text{block})}(\boldsymbol{x}) \\ &\quad \le \max\{\widehat{f}_n^{(\text{max-min})}(\boldsymbol{x}), \widehat{f}_n^{(\text{min-max})}(\boldsymbol{x})\} \quad \forall \boldsymbol{x} \in V, \end{aligned} \tag{16}$$

where $\widehat{f}_n^{(\text{max-min})}$ and $\widehat{f}_n^{(\text{min-max})}$ are the block max-min and min-max estimators given in (10). It is clear from (10) that both the max-min and min-max estimators are nondecreasing on the graph $G = (V, E)$ as the maximum is taken over increasing classes indexed by $\boldsymbol{x} \in V$ and the minimum over decreasing classes. However, the monotonicity of the block estimator,

$\widehat{f}_n^{(\mathrm{block})} \in \mathcal{F}$ or even $\widehat{\boldsymbol{f}}_n^{(\mathrm{block})} \in \mathcal{F}_n$, is optional in our analysis. A practical monotone solution is

$$\widehat{f}_n^{(\mathrm{block})}(\boldsymbol{x}) = \frac{1}{2}\bigl\{\widehat{f}_n^{(\mathrm{max\text{-}min})}(\boldsymbol{x}) + \widehat{f}_n^{(\mathrm{min\text{-}max})}(\boldsymbol{x})\bigr\} \quad \forall \boldsymbol{x} \in V. \tag{17}$$

We note that the estimator (16) is defined on the entire V. This is needed as we shall consider the L_q risk (3) as well as the ℓ_q risk (2). It would be tempting to define the block estimator by

$$\max_{\boldsymbol{u} \preceq \boldsymbol{x}} \min_{\boldsymbol{x} \preceq \boldsymbol{v}} \overline{y}_{[\boldsymbol{u},\boldsymbol{v}]} \le \widehat{f}^{(\mathrm{block})}(\boldsymbol{x}) \le \min_{\boldsymbol{x} \preceq \boldsymbol{v}} \max_{\boldsymbol{u} \preceq \boldsymbol{x}} \overline{y}_{[\boldsymbol{u},\boldsymbol{v}]}$$

(Fokianos, Leucht and Neumann (2017)). However, unfortunately, when $\boldsymbol{x}$ is not a design point, $\overline{y}_{[\boldsymbol{u},\boldsymbol{v}]}$ is undefined when $[\boldsymbol{u},\boldsymbol{v}]$ contains no data point, and $\widehat{f}_n^{(\mathrm{max\text{-}min})}(\boldsymbol{x}) \le \widehat{f}_n^{(\mathrm{min\text{-}max})}(\boldsymbol{x})$ is not guaranteed to hold even for properly defined max-min and min-max estimators in (10), even in the univariate case. For example, for $V = [0, 1]$ with two data points $(x_1, y_1) = (0, 1)$ and $(x_2, y_2) = (1, 2)$, (10) gives $\widehat{f}_n^{(\mathrm{max\text{-}min})}(0.5) = 2 > 1 = \widehat{f}_n^{(\mathrm{min\text{-}max})}(0.5)$. We do have

$$\widehat{f}_n^{(\mathrm{max\text{-}min})}(\boldsymbol{x}_i) \le \widehat{f}_n^{(\mathrm{min\text{-}max})}(\boldsymbol{x}_i), \quad i = 1, \dots, n, \tag{18}$$

but the minimax formula $\widehat{f}_n^{(\mathrm{max\text{-}min})} = \widehat{f}_n^{(\mathrm{min\text{-}max})}$ may fail even on the design points as the example in Figure 1 demonstrates.

In the rest of this section, we prove that the max-min and min-max estimators defined with upper and lower sets in a graph G, including the LSE, can always be expressed as the block estimators defined as in (16) but over a larger graph than G, so that our analysis of general block estimators is also relevant to the LSE. We present our argument in a more general setting as follows.

Formally, a subset of vertices $U \subseteq V$ is called an upper set if $U = \{\boldsymbol{x} : f(\boldsymbol{x}) > t\}$ for some $f \in \mathcal{F}$ and real t, or equivalently the indicator function 1_U is non-decreasing on G, that is, $1_U \in \mathcal{F}$; a subset $L \subseteq V$ is called a lower set if $L = \{\boldsymbol{x} : f(\boldsymbol{x}) \le t\}$ for some $f \in \mathcal{F}$ and $t \in \mathbb{R}$, that is, the complement of an upper set. Let $\mathcal{U}$ be the collection of all upper sets, $\mathcal{L}$ the collection of all lower sets, and

$$\mathcal{U}_{\boldsymbol{x}} \subseteq \{U \in \mathcal{U} : \boldsymbol{x} \in U\} \quad \text{and} \quad \mathcal{L}_{\boldsymbol{x}} \subseteq \{L \in \mathcal{L} : \boldsymbol{x} \in L\}$$

be certain subsets of the collections of upper and lower sets containing $\boldsymbol{x}$. The max-min and min-max estimator can be defined in general as

$$\begin{aligned} \widehat{f}_n^{(\mathrm{max\text{-}min})}(\boldsymbol{x}) &= \max_{U \in \mathcal{U}_{\boldsymbol{x}}, n_U > 0} \min_{L \in \mathcal{L}_{\boldsymbol{x}}, n_{U\cap L} > 0} \overline{y}_{U \cap L}, \quad \boldsymbol{x} \in V, \\ \widehat{f}_n^{(\mathrm{min\text{-}max})}(\boldsymbol{x}) &= \min_{L \in \mathcal{L}_{\boldsymbol{x}}, n_L > 0} \max_{U \in \mathcal{U}_{\boldsymbol{x}}, n_{U\cap L} > 0} \overline{y}_{U \cap L}, \quad \boldsymbol{x} \in V, \end{aligned} \tag{19}$$

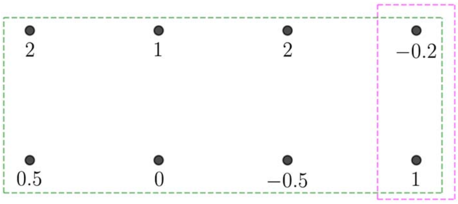


FIG. 1. *Responses y_i on a 4×2 lattice design: At design point $\boldsymbol{x} = (4, 1)$, $\widehat{f}_n^{(\mathrm{max\text{-}min})}(\boldsymbol{x}) = 0.4$ is attained by the mean inside the magenta box and $\widehat{f}_n^{(\mathrm{max\text{-}min})}(\boldsymbol{x}) = 0.725$ attained by the mean inside the green box.*

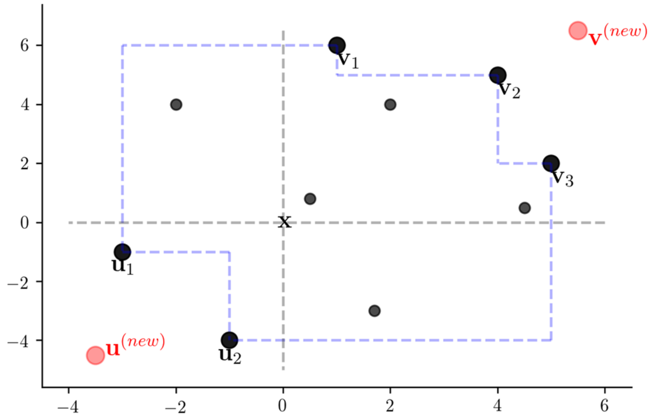


FIG. 2. *Amendment of G to include $U \cap L = \bigcup_{j\in\{1,2\},k\in\{1,2,3\}}[\boldsymbol{u}_j, \boldsymbol{v}_k]$ where $\boldsymbol{u}^{(\mathrm{new})}$ has two inbound edges from $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ and $\boldsymbol{v}^{(\mathrm{new})}$ has three outbound edges to $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$.*

where $n_A = \{i \le n : \boldsymbol{x}_i \in A\}$. These max-min and min-max estimators are nondecreasing in $\boldsymbol{x}$ on the entire graph if $\mathfrak{U}_{\boldsymbol{x}}$ is nondecreasing in $\boldsymbol{x}$ and $\mathfrak{L}_{\boldsymbol{x}}$ non-increasing in $\boldsymbol{x}$: $\mathfrak{U}_{\boldsymbol{x}} \subseteq \mathfrak{U}_{\boldsymbol{x}'}$ and $\mathfrak{L}_{\boldsymbol{x}} \supseteq \mathfrak{L}_{\boldsymbol{x}'}$ for all ordered pairs $\boldsymbol{x} \preceq \boldsymbol{x}'$.

By (9), the LSE is a special case of (19) when $\mathfrak{U}_{\boldsymbol{x}}$ and $\mathfrak{L}_{\boldsymbol{x}}$ are taken to be the largest possible. The block max-min and min-max estimators (10) are special cases of (19) with $\mathfrak{U}_{\boldsymbol{x}} = \{[\boldsymbol{u}, *] : \boldsymbol{u} \preceq \boldsymbol{x}\}$ and $\mathfrak{L}_{\boldsymbol{x}} = \{[*, \boldsymbol{v}] : \boldsymbol{x} \preceq \boldsymbol{v}\}$. Conversely, the LSE, and more generally (19), can be written as

$$\begin{aligned}\widehat{f}_n^{(\text{max-min})}(\boldsymbol{x}) &= \max_{\boldsymbol{u}\in A_{\boldsymbol{x}}, n_{\boldsymbol{u},*}>0}\ \min_{\boldsymbol{v}\in B_{\boldsymbol{x}}, n_{\boldsymbol{u},\boldsymbol{v}}>0} \overline{y}_{[\boldsymbol{u},\boldsymbol{v}]}, \quad \boldsymbol{x} \in V,\\ \widehat{f}_n^{(\text{min-max})}(\boldsymbol{x}) &= \min_{\boldsymbol{v}\in B_{\boldsymbol{x}}, n_{*,\boldsymbol{v}}>0}\ \max_{\boldsymbol{u}\in A_{\boldsymbol{x}}, n_{\boldsymbol{u},\boldsymbol{v}}>0} \overline{y}_{[\boldsymbol{u},\boldsymbol{v}]}, \quad \boldsymbol{x} \in V,\end{aligned} \tag{20}$$

based on the average response in blocks $[\boldsymbol{u}, \boldsymbol{v}]$ for suitable $A_{\boldsymbol{x}}$ and $B_{\boldsymbol{x}}$ in a larger graph G^* in which G is a subgraph. We define G^* by amending G with new nodes and edges as follows. For each upper set U, we amend G with node $\boldsymbol{u}^{(\mathrm{new})} = \boldsymbol{u}^{(\mathrm{new},U)}$ and edges $\{\boldsymbol{u} \to \boldsymbol{u}^{(\mathrm{new})} : \boldsymbol{u} \in U\}$, whereas for each lower set L, we amend G with node $\boldsymbol{v}^{(\mathrm{new})} = \boldsymbol{v}^{(\mathrm{new},L)}$ and edges $\{\boldsymbol{v}^{(\mathrm{new})} \to \boldsymbol{v} : \boldsymbol{v} \in L\}$. Define in the new graph G^* the estimators (20) with $A_{\boldsymbol{x}} = \{\boldsymbol{u}^{(\mathrm{new},U)} : U \in \mathfrak{U}_{\boldsymbol{x}}\}$ and $B_{\boldsymbol{x}} = \{\boldsymbol{v}^{(\mathrm{new},L)} : L \in \mathfrak{L}_{\boldsymbol{x}}\}$. Then the restriction of (20) on G is identical to (19) as $[\boldsymbol{u}^{(\mathrm{new},U)}, \boldsymbol{v}^{(\mathrm{new},L)}]$ contains the same set of design points as $U \cap L$. This can be seen as follows. For any pair of upper and lower sets U and L, $[\boldsymbol{u}^{(\mathrm{new},U)}, \boldsymbol{v}^{(\mathrm{new},L)}] \supset U \cap L$ by the definition of $\boldsymbol{u}^{(\mathrm{new},U)}$ and $\boldsymbol{v}^{(\mathrm{new},L)}$ and the associated collections of new edges. On the other hand, for any design point $\boldsymbol{x}_i \in [\boldsymbol{u}^{(\mathrm{new},U)}, \boldsymbol{v}^{(\mathrm{new},L)}]$, $\boldsymbol{u}^{(\mathrm{new},U)} \preceq \boldsymbol{x}_i$ could happen only if $\boldsymbol{u} \preceq \boldsymbol{x}_i$ for some $\boldsymbol{u} \in U$ as there is no other way to connect to $\boldsymbol{u}^{(\mathrm{new},U)}$ in G^*, while $\boldsymbol{x}_i \preceq \boldsymbol{v}^{(\mathrm{new},L)}$ could happen only if $\boldsymbol{x}_i \preceq \boldsymbol{v}$ for some $\boldsymbol{v} \in L$. Thus, $\overline{y}_{U\cap L} = \overline{y}_{[\boldsymbol{u}^{(\mathrm{new},U)}, \boldsymbol{v}^{(\mathrm{new},L)}]}$. Figure 2 demonstrate a $[\boldsymbol{u}^{(\mathrm{new})}, \boldsymbol{v}^{(\mathrm{new})}]$ when G is a 2-dimensional lattice.

Our theoretical results on general graph in Section 3.1 below are applicable to the LSE by writing the LSE as a block estimator on a much larger amended graph. However, the more specific results in multiple isotonic regression in Sections 3.2–3.7 are not application to the LSE as they are based the calculation of the variability bounds in (21) and (22) below for the lattice and random designs, not on the enlarged graph.

3. Theoretical results.

In this section, we first analyze the block estimator $\widehat{f}_n^{(\mathrm{block})}(\boldsymbol{x})$ in (16) for graphs under the most general setting. Specific risk bounds are then given for multiple isotonic regression with fixed lattice designs and random designs.

3.1. *General isotonic regression on graph.* We shall extend the risk bounds of Zhang (2002) from the real line to general graphs. To this end, we first derive an upper bound for the total risk in subsets $V_0 \subset V$,

$$T_q(V_0) = \sum_{\boldsymbol{x}_i \in V_0} \mathbb{E}\bigl|\widehat{f}_n^{(\text{block})}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\bigr|^q,$$

based on the value of the true f on V_0. Such bounds automatically produce adaptive risk bounds when the true f is "piecewise constant" in a partition of V. Given V_0, let $r_{q,+}(m)$ be a nonincreasing function of $m \in \mathbb{N}^+$ satisfying

$$r_{q,+}(m) \ge \max\biggl\{\mathbb{E}\biggl(\max_{\boldsymbol{u} \preceq \boldsymbol{x}} \sum_{\boldsymbol{x}_i \in [\boldsymbol{u},\boldsymbol{v}]} \frac{\varepsilon_i}{n_{\boldsymbol{u},\boldsymbol{v}}}\biggr)_+^q : n_{\boldsymbol{x},\boldsymbol{v}} = m, \boldsymbol{x} \preceq \boldsymbol{v} \text{ and } \boldsymbol{v} \in V_0\biggr\}. \tag{21}$$

This function bounds the error of the block estimator from the positive side when the positive part of its bias is no greater than the positive part of the maximum average of at least m noise variables. Similarly, to control the estimation error from the negative side, let $r_{q,-}(m)$ be a nonincreasing function satisfying

$$r_{q,-}(m) \ge \max\biggl\{\mathbb{E}\biggl(\min_{\boldsymbol{v} \succeq \boldsymbol{x}} \sum_{\boldsymbol{x}_i \in [\boldsymbol{u},\boldsymbol{v}]} \frac{\varepsilon_i}{n_{\boldsymbol{u},\boldsymbol{v}}}\biggr)_-^q : n_{\boldsymbol{u},\boldsymbol{x}} = m, \boldsymbol{u} \preceq \boldsymbol{x} \text{ and } \boldsymbol{u} \in V_0\biggr\}. \tag{22}$$

With the above functions $r_{q,\pm}(m)$, we define for $\boldsymbol{x} \in V_0$,

$$\begin{aligned}
m_{\boldsymbol{x},-} &= \max\bigl\{n_{\boldsymbol{u},\boldsymbol{x}} : f(\boldsymbol{u}) \ge f(\boldsymbol{x}) - r_{q,-}^{1/q}(n_{\boldsymbol{u},\boldsymbol{x}}), \boldsymbol{u} \preceq \boldsymbol{x} \text{ and } \boldsymbol{u} \in V_0\bigr\},\\
\boldsymbol{u}_{\boldsymbol{x}} &= \mathop{\arg\max}_{\boldsymbol{u} \in V_0 : \boldsymbol{u} \preceq \boldsymbol{x}} \bigl\{n_{\boldsymbol{u},\boldsymbol{x}} : f(\boldsymbol{u}) \ge f(\boldsymbol{x}) - r_{q,-}^{1/q}(n_{\boldsymbol{u},\boldsymbol{x}})\bigr\},\\
m_{\boldsymbol{x}} = m_{\boldsymbol{x},+} &= \max\bigl\{n_{\boldsymbol{x},\boldsymbol{v}} : f(\boldsymbol{v}) \le f(\boldsymbol{x}) + r_{q,+}^{1/q}(n_{\boldsymbol{x},\boldsymbol{v}}), \boldsymbol{x} \preceq \boldsymbol{v} \text{ and } \boldsymbol{v} \in V_0\bigr\},\\
\boldsymbol{v}_{\boldsymbol{x}} &= \mathop{\arg\max}_{\boldsymbol{v} \in V_0 : \boldsymbol{x} \preceq \boldsymbol{v}} \bigl\{n_{\boldsymbol{x},\boldsymbol{v}} : f(\boldsymbol{v}) \le f(\boldsymbol{x}) + r_{q,+}^{1/q}(n_{\boldsymbol{x},\boldsymbol{v}})\bigr\}.
\end{aligned} \tag{23}$$

Roughly speaking, the above quantities provide configurations in which the bias of $\widehat{f}_n(\boldsymbol{x}_i)$ is of no greater order than its variability from the negative and positive sides, so that the error of the block estimator is of no greater order than an average of $m_{\boldsymbol{x}_i,-}$ noise variables on the negative side and the average of $m_{\boldsymbol{x}} = m_{\boldsymbol{x}_i,+}$ noise variables on the positive side. Thus, it makes sense to count the frequencies of $m_{\boldsymbol{x}_i,-}$ and $m_{\boldsymbol{x}_i}$ as follows:

$$\begin{aligned}
\ell_-(m) &= \#\{i : \boldsymbol{x}_i \in V_0, m_{\boldsymbol{x}_i,-} \le m\},\\
\ell_+(m) &= \#\{i : \boldsymbol{x}_i \in V_0, m_{\boldsymbol{x}_i} \le m\}.
\end{aligned} \tag{24}$$

We note that the functions $r_{q,\pm}$ in (21) and (22) do not depend on f, and all the quantities in (23) and (24) depend on the true f only through $\{f(\boldsymbol{x}) : \boldsymbol{x} \in V_0\}$.

THEOREM 1. *Assume f is nondecreasing on a graph $G = (V, E)$. Let $r_{q,\pm}(m)$ be given by* (21) *and* (22), *and $\ell_\pm(m)$ by* (24). *Then it holds for any block estimator $\widehat{f}_n^{(\text{block})}(\boldsymbol{x})$ in* (16) *that*

$$\begin{aligned}
\mathbb{E}\bigl\{\widehat{f}_n^{(\text{block})}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\bigr\}_+^q &\le 2^q r_{q,+}(m_{\boldsymbol{x}_i}) \quad \forall \boldsymbol{x}_i \in V_0,\\
\mathbb{E}\bigl\{\widehat{f}_n^{(\text{block})}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\bigr\}_-^q &\le 2^q r_{q,-}(m_{\boldsymbol{x}_i,-}) \quad \forall \boldsymbol{x}_i \in V_0.
\end{aligned} \tag{25}$$

Consequently, for any upper bounds $\ell^*_\pm(m) \ge \ell_\pm(m)$ *with* $\ell^*_\pm(0) = 0$,

$$T_q(V_0) \le \sum_{m=1}^{\infty} 2^q r_{q,+}(m)\{\ell^*_+(m) - \ell^*_+(m-1)\} \tag{26}$$
$$+ \sum_{m=1}^{\infty} 2^q r_{q,-}(m)\{\ell^*_-(m) - \ell^*_-(m-1)\}.$$

Theorem 1 provides risk bound for the block estimator (16) over a subset V_0 of design points in terms of upper bound functions $r_{q,\pm}(m)$ and $\ell^*_\pm(m)$. Ideally, we would like to have

$$r_{q,\pm}(m) = C_{q,d}\sigma^q m^{-q/2} \tag{27}$$

in (21) and (22). When the design points in V_0 are linear and the $(q \vee 2)$th moment of the noise variable is uniformly bounded, (21) and (22) hold for the above choice of $r_{q,\pm}(m)$. This choice of $r_{q,\pm}(m)$ is also valid when V is a lattice in $\mathbb{R}^d$ and ε_i are independent variables with uniformly bounded $(q \vee 2)$th moment, as we will prove in Section 3.3.

REMARK 1. Suppose the nondecreasing function f satisfies extra constraints, that is, f belongs to a subclass $\mathcal{F}_0 \subset \mathcal{F}$. We may use the block estimators to compute a nondecreasing solution and then project it to the subclass $\mathcal{F}_0$. This two-step estimator must produce loss no greater than that of the block estimators whenever $\mathcal{F}_0$ is convex and at most twice the loss for general $\mathcal{F}_0$. Consequently, the risk bound in Theorem 1 and all others produced later in this paper (Theorems 2-9) remain valid for the two-step estimator. This set of results may serve as a benchmark for estimation under constraints more than just monotonicity.

3.2. *Minimax lower bound in multiple isotonic regression with lattice designs.* We study in the rest of this section multiple isotonic regression in $V \subseteq \mathbb{R}^d$ where $\boldsymbol{a} \preceq \boldsymbol{b}$ iff $\boldsymbol{a} \le \boldsymbol{b}$, that is, $a_j \le b_j \forall 1 \le j \le d$, for all $\boldsymbol{a} = (a_1, \dots, a_d)^T$ and $\boldsymbol{b} = (b_1, \dots, b_d)^T$, and $\mathfrak{F}$ is the class of all nondecreasing functions $f(t_1, \dots, t_d) \uparrow t_j, \forall j = 1, \dots, d$.

The lattice design we are considering is given by

$$V = \{\boldsymbol{x}_i : 1 \le i \le n\} = [\mathbf{1}, \boldsymbol{n}] = \prod_{j=1}^{d} \{1, \dots, n_j\}, \tag{28}$$

where $\boldsymbol{n} = (n_1, \dots, n_d)^T$ with positive integers n_j and $n = \prod_{j=1}^d n_j$. Here, $[\mathbf{1}, \boldsymbol{n}]$ is treated as a set of integer-valued vectors in $\mathbb{N}^d$, forming a lattice. Occasionally, we may also use $[\boldsymbol{u}, \boldsymbol{v}]$ to denote a hyperrectangle of real numbers in continuum. This slight abuse of notation typically would not lead to confusion, for example, in $\boldsymbol{x}_i \in [\boldsymbol{u}, \boldsymbol{v}]$, but we would be specific if necessary. Without loss of generality, we assume in this subsection $n_1 \ge n_2 \ge \cdots \ge n_d$. In the above lattice design, we provide a minimax lower bound in multiple isotonic regression as follows.

PROPOSITION 1. *Suppose* $\varepsilon_i \sim N(0, \sigma^2)$. *Let* $\Delta(\boldsymbol{f}_n/\sigma) = \{f(\boldsymbol{n}) - f(\mathbf{1})\}/\sigma$, $n_{d+1} = 1$, $n^*_s = \prod_{j=1}^s n_j$, $t_s = n^*_s/n^s_s$, $t_{d+2} = \infty$ *and* $s_q = \lceil 2/(q-1) \rceil \wedge (d+1)$. *Let* $h_0(t) = \Delta^*_n \sqrt{t}$ *and define piecewise* $H(t) = \min\{1, h_0(t)/(n^*_s/t)^{1/(s \wedge d)}\}$, $t \in [t_s, t_{s+1}]$, $s = 1, \dots, d+1$. *Then*

$$\inf_{\widehat{\boldsymbol{f}}} \sup\{R_q(\widehat{\boldsymbol{f}}, \boldsymbol{f}_n) : \boldsymbol{f}_n \in \mathfrak{F}_n, \Delta(\boldsymbol{f}_n/\sigma) \le \Delta^*_n\}$$
$$\gtrsim_{q,d} \sigma^q \max\{(t \wedge n)^{-q/2} H(t) : t \wedge h_0(t) \ge 1\} \tag{29}$$

$$= \sigma^q \times \begin{cases} 1, & n_1 \le \Delta_n^*, & (s=0) \\ \big(\Delta_n^*/(n_s^*)^{1/s}\big)^{\frac{qs}{2+s}} & n_{s+1}/t_{s+1}^{1/2} \le \Delta_n^* \le n_s/t_s^{1/2}, & (1 \le s < s_q) \\ \Delta_n^*/\big(n_s t_s^{(q-1)/2}\big), & t_s^{-1/2} \le \Delta_n^* \le n_s/t_s^{1/2}, & (s = s_q \le d) \\ (\Delta_n^*)^{q-2/s}/(n_s^*)^{1/s}, & t_{s+1}^{-1/2} \le \Delta_n^* \le t_s^{-1/2}, & (s_q \le s \le d) \\ n^{-q/2}, & 0 \le \Delta_n^* \le n^{-1/2}. & (s = d+1) \end{cases}$$

In particular, when $n_1 = \cdots = n_d = n^{1/d}$ and $\Delta_n^* \ge n^{-1/2}$, the right-hand side of (29) *is*

$$\sigma^q \times \begin{cases} \min\{1, (\Delta_n^*/n^{1/d})^{qd/(d+2)}\}, & q \le 1 + 2/d, \\ \min\{1, \Delta_n^*/n^{1/d}, (\Delta_n^*)^{q-2/d}/n^{1/d}\}, & q \ge 1 + 2/d. \end{cases} \tag{30}$$

On the right-hand side of (29), the breaking points on $[0, \infty)$ for Δ_n^* are

$$0, n^{-1/2} = t_{d+1}^{-1/2}, t_d^{-1/2}, \dots, t_{s_q}^{-1/2}, n_{s_q}/t_{s_q}^{1/2}, \dots, n_1/t_1^{1/2} = n_1.$$

Note that 1 lies in between $t_{s_q}^{-1/2}$ and $n_{s_q}/t_{s_q}^{1/2}$. The above minimax lower bound also depends on the loss function through q and the dimension of the lattice. For $q \ge 3$, we have $s_q = 1$, so that

$$\inf_{\widehat{\boldsymbol{f}}} \sup\{R_q(\widehat{\boldsymbol{f}}, \boldsymbol{f}_n) : \boldsymbol{f}_n \in \mathcal{F}_n, \Delta(\boldsymbol{f}_n/\sigma) \le \Delta_n^*\} \gtrsim_{q,d} \sigma^q \min(1, \Delta_n^*/n_1)$$

for $\Delta_n^* \ge 1$. However for $q = 2$, we have $s_q = 2$, so that (29) yields

$$\begin{aligned} &\inf_{\widehat{\boldsymbol{f}}} \sup\{R_2(\widehat{\boldsymbol{f}}, \boldsymbol{f}_n) : \boldsymbol{f}_n \in \mathcal{F}_n, \Delta(\boldsymbol{f}_n/\sigma) \le \Delta_n^*\} \\ &\gtrsim_d \sigma^2 \times \begin{cases} 1, & n_1 \le \Delta_n^*, & (s=0) \\ (\Delta_n^*/n_1)^{2/3}, & n_2^{3/2}/n_1^{1/2} \le \Delta_n^* \le n_1, & (s=1) \\ \Delta_n^*/(n_1 n_2)^{1/2}, & \sqrt{n_2/n_1} \le \Delta_n^* \le n_2^{3/2}/n_1^{1/2}. & (s=2) \end{cases} \end{aligned} \tag{31}$$

For $\Delta_n^* \asymp 1$, this matches the lower bound for the ℓ_2 minimax rate in Chatterjee, Guntuboyina and Sen (2018) for $d = 2$ and Han et al. (2019) for $d \ge 3$. For $5/3 \le q < 2 \le d$, we have $s_q = 3$.

If (29) is achievable, the integer parameter s can be viewed as the effective dimension of the isotonic regression problem as the rate depends on $\boldsymbol{n}$ only through $n_1, \dots, n_s$ when n_{s+1} is sufficiently small; the rate would also be achievable by separate s-dimensional isotonic regression in the $\prod_{j=s+1}^d n_j = n/n_s^*$ individual s-dimensional sheets with fixed $x_{s+1}, \dots, x_d$. For example, in (31), the minimax rate can be achieved by $\widehat{\boldsymbol{f}}_n = \boldsymbol{y}$ for $s = 0$, by the row-by-row univariate isotonic regression for $s = 1$, and by individual bivariate isotonic least squares up to a factor of $(\log n)^4$ for $s = 2$ (Chatterjee, Guntuboyina and Sen (2018)). We will prove in the next subsection that the block estimator (16) achieves the rate in (29) for a wide range of Δ_n^*, so that Proposition 1 indeed provides the minimax rate.

In the proof of Proposition 1, we divide $[\boldsymbol{1}, \boldsymbol{n}'] \subset V = [\boldsymbol{1}, \boldsymbol{n}]$ into a $K_1 \times \cdots \times K_d$ lattice of hyperrectangles of size $m_1 \times \cdots \times m_d$, indexed by $\boldsymbol{k} = (k_1, \dots, k_d)^T$, $k_j = 1, \dots, K_j$, $j = 1, \dots, d$, and consider the class of piecewise constant functions $f(\boldsymbol{x}) = g(\boldsymbol{k})$ satisfying

$$g(\boldsymbol{k}) = \sigma \min\{\Delta_n^*, (m^*)^{-1/2}[\theta(\boldsymbol{k}) + (k_1 + \cdots + k_d - k^*)_+]\}, \quad \theta(\boldsymbol{k}) \in \{0, 1\},$$

and $f(\boldsymbol{x}) = \sigma \Delta_n^*$ for $\boldsymbol{x} \in [\boldsymbol{1}, \boldsymbol{n}] \setminus [\boldsymbol{1}, \boldsymbol{n}']$, where $m^* = \prod_{j=1}^d m_j$ is the size of the hyperrectangle. As $g(\boldsymbol{k})$ is nondecreasing in k_j for each j for all $\theta(\boldsymbol{k}) \in \{0, 1\}$, this construction provides a lower bound for the ℓ_q risk proportional to the product of $\sigma^q (m^*)^{-q/2}$ and the number of free $\theta(\boldsymbol{k})$. This is summarized in the following lemma.

LEMMA 1. *Under the conditions of Proposition* 1,

$$\begin{aligned}&\inf_{\widehat{\boldsymbol{f}}} \sup\{\mathbb{E}\|\widehat{\boldsymbol{f}} - \boldsymbol{f}_n\|_q^q : \boldsymbol{f}_n \in \mathcal{F}_n, \Delta(\boldsymbol{f}_n/\sigma) \le \Delta_n^*\}\\ &\quad \ge c_q c_d \sigma^q n \max_{\boldsymbol{m}\in\mathcal{M}} \left\{ \frac{1}{(m^*)^{q/2}} \min\left(\frac{\sqrt{m^*}\Delta_n^*}{\max_j \lfloor n_j/m_j \rfloor}, 1 \right) \right\},\end{aligned} \tag{32}$$

where $c_q = \inf_\delta \mathbb{E}_{\mu\sim\text{Bernoulli}(1/2)} |\delta(N(\mu,1)) - \mu|_q^q$ *is the Bayes risk for estimating* μ *with the Bernoulli*$(1/2)$ *prior based on a single* $N(\mu,1)$ *observation*, c_d *is a constant depending on* d *only*,

$$\mathcal{M} = \{\boldsymbol{m} = (m_1, \dots, m_d) : m_j \in \mathbb{N}_+, m_j \le n_j \ \forall j \le d, \sqrt{m^*}\Delta_n^* \ge 1\},$$

and $m^* = \prod_{j \le d} m_j$. *Moreover, the optimal configuration of* $\boldsymbol{m}$ *in* (32) *must satisfy either* $m_j = 1$ *or* $\lfloor n_j/m_j \rfloor = \max_{1\le j\le d} \lfloor n_j/m_j \rfloor$ *for each* j.

3.3. *The block estimator in multiple isotonic regression with lattice designs.* We further divide this subsection into three separate sub-subsections to study the performance of the block estimator at a single design point $\boldsymbol{x}_i$, in an arbitrary subblock $[\boldsymbol{a}, \boldsymbol{b}] \subset [\boldsymbol{1}, \boldsymbol{n}]$, and on the entire lattice $[\boldsymbol{1}, \boldsymbol{n}]$. It is of great interest to show that the block estimator in (16) matches the minimax lower bound given in Proposition 1, which will be done in the third sub-subsection for general q and d.

3.3.1. *Risk of the block estimator at a single design point.* For any given point in the design lattice, the following proposition asserts that the block estimator matches certain one-sided oracle estimators in the rate of one-sided L_q risks.

PROPOSITION 2. *Let* $\widehat{f}_n^{(\text{block})}(\boldsymbol{x})$ *be the block estimator in* (16) *with the lattice design* $V = [\boldsymbol{1}, \boldsymbol{n}]$ *in* (28). *Let* $q \ge 1$ *and* $r_{q,\pm}(m)$ *be as in* (21) *and* (22). *Assume* ε_i *are independent* $N(0,\sigma^2)$ *random variables. Then, for any design point* $\boldsymbol{x}_i \in [\boldsymbol{1}, \boldsymbol{n}]$,

$$\mathbb{E}(\widehat{f}_n^{(\text{block})}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i))_+^q \le 2^q r_{q,+}(m_{\boldsymbol{x}_i}) \le C_{q,d} \min_{\boldsymbol{x}_i \le \boldsymbol{v} \le \boldsymbol{n}} \mathbb{E}(\overline{y}_{[\boldsymbol{x}_i,\boldsymbol{v}]} - f(\boldsymbol{x}_i))_+^q, \tag{33}$$

where $\overline{y}_{[\boldsymbol{u},\boldsymbol{v}]} = \sum_{\boldsymbol{u}\le\boldsymbol{x}_i\le\boldsymbol{v}} y_i/n_{\boldsymbol{u},\boldsymbol{v}}$, *and*

$$\mathbb{E}(\widehat{f}_n^{(\text{block})}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i))_-^q \le 2^q r_{q,-}(m_{\boldsymbol{x}_i}) \le C_{q,d} \min_{\boldsymbol{1} \le \boldsymbol{u} \le \boldsymbol{x}_i} \mathbb{E}(\overline{y}_{[\boldsymbol{u},\boldsymbol{x}_i]} - f(\boldsymbol{x}_i))_-^q. \tag{34}$$

Consequently, with $\mathbb{E}_g$ *being the expectation under which* $y_i = g(\boldsymbol{x}_i) + \varepsilon_i$,

$$\begin{aligned}&\mathbb{E}\big|\widehat{f}_n^{(\text{block})}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\big|^q\\ &\quad \le C_{q,d} \min_{\boldsymbol{u}\le\boldsymbol{x}_i\le\boldsymbol{v}} \{\mathbb{E}_g|\overline{y}_{[\boldsymbol{u},\boldsymbol{v}]} - g(\boldsymbol{x}_i)|^q : g \in \mathcal{F}, g(\boldsymbol{v}) = f(\boldsymbol{v}) \ \forall \boldsymbol{v} \ge \boldsymbol{x}_i\}\\ &\qquad + C_{q,d} \min_{\boldsymbol{u}\le\boldsymbol{x}_i\le\boldsymbol{v}} \{\mathbb{E}_g|\overline{y}_{[\boldsymbol{u},\boldsymbol{v}]} - g(\boldsymbol{x}_i)|^q : g \in \mathcal{F}, g(\boldsymbol{u}) = f(\boldsymbol{u}) \ \forall \boldsymbol{u} \le \boldsymbol{x}_i\}.\end{aligned} \tag{35}$$

Suppose we are confined to consider only block mean estimators $\overline{y}_{[\boldsymbol{u},\boldsymbol{v}]}$ with no negative bias in the estimation of $f(\boldsymbol{x}_i)$ but we also want to control the positive side of the error. As f is nondecreasing but otherwise unknown, we are thus forced to choose $\boldsymbol{u} \ge \boldsymbol{x}_i$. As $\overline{y}_{[\boldsymbol{u},\boldsymbol{v}]}$

with $\boldsymbol{x}_i \le \boldsymbol{u} \le \boldsymbol{v}$ would have larger bias and variance than $\overline{y}_{[\boldsymbol{x}_i,\boldsymbol{v}]}$, the optimal $[\boldsymbol{u},\boldsymbol{v}]$ is given by

$$\min_{\boldsymbol{u}=\boldsymbol{x}_i\le\boldsymbol{v}\le\boldsymbol{n}} \mathbb{E}\big(\overline{y}_{[\boldsymbol{x}_i,\boldsymbol{v}]} - f(\boldsymbol{x}_i)\big)_+^q.$$

The above minimum can be viewed as an oracle benchmark under the no-negative-bias constraint as the solution of the optimal $\boldsymbol{v}$ still depends on f. Although the block estimator (16) is unlikely to be unbiased, (33) and (34) assert that its one-sided risks match the rates of such oracle benchmarks from both the positive and negative sides. Another interpretation of the performance of the block estimator is (35) in which the oracle expert has to guard against the worst case scenarios in the uncertainty of f on either sides, but not simultaneously on both.

We prove Proposition 2 with an application of Theorem 1. This requires more explicit variability bounds $r_{q,\pm}(m)$ in (21) and (22) as in (27). This validity of (27) is a consequence of the following lemma, which extends Doob's inequality to certain multiple indexed submartingales. It plays a key role in removing the normality assumption on the noise $\varepsilon_1,\dots,\varepsilon_n$ in our analysis.

LEMMA 2. *Let $\mathcal{T}=\mathcal{T}_1\times\cdots\times\mathcal{T}_d\subseteq\mathbb{R}^d$ be an index set with $\mathcal{T}_j\subseteq\mathbb{R}$. Let $\{f_{\boldsymbol{t}},\boldsymbol{t}\in\mathcal{T}\}$ be a collection of random variables. Suppose for each j and each $(s_1,\dots,s_{j-1},t_{j+1},\dots,t_d)$, $\{f_{s_1,\dots,s_{j-1},t,t_{j+1},\dots,t_d}, t\in\mathcal{T}_j\}$ is a submartingale with respect to certain filtration $\{\mathcal{F}_t^{(j)}, t\in\mathcal{T}_j\}$. Then, for all $q>1$ and $\boldsymbol{t}\in\mathcal{T}$,*

$$\mathbb{E}\max_{\boldsymbol{s}\in\mathcal{T},\boldsymbol{s}\le\boldsymbol{t}} |f_{\boldsymbol{s}}|^q \le \big(q/(q-1)\big)^{qd}\mathbb{E}|f_{\boldsymbol{t}}|^q.$$

In particular when ε_i's are independent random variables with $\mathbb{E}\varepsilon_i=0$,

$$\mathbb{E}\max_{\boldsymbol{s}\le\boldsymbol{t}}\bigg|\sum_{\boldsymbol{x}_i\le\boldsymbol{s}}\varepsilon_i\bigg|^q \le \begin{cases} \big(q/(q-1)\big)^{qd}\mathbb{E}\bigg|\displaystyle\sum_{\boldsymbol{x}_i\le\boldsymbol{t}}\varepsilon_i\bigg|^q, & q\ge 2,\\ \bigg(4^d\mathbb{E}\bigg|\displaystyle\sum_{\boldsymbol{x}_i\le\boldsymbol{t}}\varepsilon_i\bigg|^2\bigg)^{q/2}, & 1\le q<2.\end{cases}$$

3.3.2. *Risk of the block estimator in a sub-block.* To automatically deal with adaptation which gives better risk bound when $f(\cdot)$ is piecewise constant, we first consider the risk in one of such "piece," a hyperrectangle $[\boldsymbol{a},\boldsymbol{b}]\subseteq V=[\boldsymbol{1},\boldsymbol{n}]$.

THEOREM 2. *Let $\widehat{f}_n^{(\text{block})}(\boldsymbol{x})$ be the block estimator in (16) with the lattice design $V=[\boldsymbol{1},\boldsymbol{n}]$ in (28). Assume ε_i are independent random variables with $\mathbb{E}\varepsilon_i=0$ and $\mathbb{E}|\varepsilon_i|^{q\vee 2}\le\sigma^{q\vee 2}$. Let $\boldsymbol{a}\le\boldsymbol{b}$ be integer vectors in $V=[\boldsymbol{1},\boldsymbol{n}]$ and $\widetilde{n}_j=b_j-a_j+1$. Suppose $\widetilde{n}_1\ge\cdots\ge\widetilde{n}_d$. Define $\widetilde{n}=n_{\boldsymbol{a},\boldsymbol{b}}$, $\widetilde{n}_{d+1}=1$, $\widetilde{n}_s^*=\prod_{j=1}^s\widetilde{n}_j$ and $t_s=\widetilde{n}_s^*/\widetilde{n}_s^s$ (with $1=t_1\le\cdots\le t_d\le t_{d+1}=\widetilde{n}$). Then, for $q\ge1$ and any $f\in\mathcal{F}$ with $\Delta_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{f}_n/\sigma)=\{f(\boldsymbol{b})-f(\boldsymbol{a})\}/\sigma\le\Delta_n^*$,*

$$\begin{aligned} T_q([\boldsymbol{a},\boldsymbol{b}]) &= \sum_{\boldsymbol{x}_i\in[\boldsymbol{a},\boldsymbol{b}]}\mathbb{E}\big|\widehat{f}_n^{(\text{block})}(\boldsymbol{x}_i)-f(\boldsymbol{x}_i)\big|^q\\ &\le C_{q,d}^* n_{\boldsymbol{a},\boldsymbol{b}}\sigma^q\bigg(\widetilde{H}(1)+\int_1^{n_{\boldsymbol{a},\boldsymbol{b}}}\frac{\widetilde{H}(dt)}{t^{q/2}}+\frac{1}{n_{\boldsymbol{a},\boldsymbol{b}}}\prod_{j=1}^d\int_0^{\widetilde{n}_j}\frac{dt}{(t\vee1)^{q/2}}\bigg),\end{aligned} \tag{36}$$

where $\widetilde{H}(t)$ is a nondecreasing and continuous function of t, defined piecewise by $\widetilde{H}(t)=\min\{1,\Delta_n^* t^{1/2}(t/\widetilde{n}_s^*)^{1/s}\}$ for $t_s\le t\le t_{s+1}$, $s=1,\dots,d$, and $C_{q,d}^*$ is continuous in $q\in[1,\infty)$

and nondecreasing in d. Moreover,

$$\begin{aligned}&\widetilde{H}(1)+\int_1^{n_{\boldsymbol{a},\boldsymbol{b}}} t^{-q/2}\widetilde{H}(dt)\\ &\lesssim_{q,d}\begin{cases}1, & \widetilde{n}_1\le\Delta_n^*, & (s=0)\\ \big(\Delta_n^*/(\widetilde{n}_s^*)^{1/s}\big)^{qs/(2+s)}, & \widetilde{n}_{s+1}/t_{s+1}^{1/2}\le\Delta_n^*\le\widetilde{n}_s/t_s^{1/2}, & (1\le s<s_q)\\ \big(\Delta_n^*/(\widetilde{n}_s t_s^{(q-1)/2})\big)\Lambda_s, & \Delta_n^*\le\widetilde{n}_s/t_s^{1/2}, & (s=s_q\le d)\end{cases}\end{aligned} \tag{37}$$

where $s_q=\lceil 2/(q-1)\rceil\wedge(d+1)$ is as in Proposition 1 and

$$\Lambda_s=\bigg[\log_+\bigg(\min\bigg\{\frac{\widetilde{n}_s}{\widetilde{n}_{s+1}},\frac{\widetilde{n}_s/(\widetilde{n}_s^*)^{1/(s+2)}}{(\Delta_n^*)^{2/(s+2)}}\bigg\}\bigg)\bigg]^{I\{2/(q-1)=s\}}. \tag{38}$$

REMARK 2. The last component on the right-hand side of (36) is bounded by

$$\sigma^q\prod_{j=1}^d\int_0^{b_j-a_j+1}\frac{dt}{(t\vee 1)^{q/2}}\lesssim_{q,d}\sigma^q\bigg[n_{\boldsymbol{a},\boldsymbol{b}}^{1-q/2}+\bigg(\prod_{j=1}^d\log_+(b_j-a_j+1)\bigg)^{I\{q=2\}}\bigg]. \tag{39}$$

When $\Delta_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{f}_n/\sigma)=0$, $\widetilde{H}(t)=0$ for all t, so that (39) is an upper bound for the rate of the total risk $T_q([\boldsymbol{a},\boldsymbol{b}])$ in the block $[\boldsymbol{a},\boldsymbol{b}]$ by Theorem 2, for any $\boldsymbol{a}\le\boldsymbol{b}$. This yields the adaptation rate stated in Section 3.4.

REMARK 3. The function $\widetilde{H}(t)$ is defined in the same way as $H(t)$ is in Proposition 1 but for the dimensions $\{\widetilde{n}_j=b_j-a_j+1, j\le d\}$ of $[\boldsymbol{a},\boldsymbol{b}]$ and range-to-noise ratio within $[\boldsymbol{a},\boldsymbol{b}]$. When $[\boldsymbol{a},\boldsymbol{b}]=[\boldsymbol{1},\boldsymbol{n}]$, we have $\widetilde{H}(t)=H(t)$ for all $t\in[1,n]$. Thus, as discussed below (31), the integer parameter s in (37), completely determined by $\{\widetilde{n}_j\}$, Δ_n^* and q, has the interpretation as the effective dimension for the estimation of f in $[\boldsymbol{a},\boldsymbol{b}]$ subject to $\{f(\boldsymbol{b})-f(\boldsymbol{a})\}/\sigma\le\Delta_n^*$. We note that as $\widetilde{H}(t)$ is a smooth fit of pieces proportional to $t^{1/2+1/s}$ or 1, the upper limit of the integration is actually $t_*=\min\{t\ge 1:\widetilde{H}(t)=1 \text{ or } t=\widetilde{n}_{\boldsymbol{a},\boldsymbol{b}}\}$, which depends on Δ_n^*, and the effective dimension s is then determined by the comparison between t_* and t_s and the critical s_q.

In addition to the validity of (27) as variability bounds in (21) and (22), which follows from Lemme 2, the proof of Theorem 2 requires the complexity bounds for the $\ell_\pm(m)$ in (24). We outline here an analysis of the count $\ell_+(m)$ in (24) in the case where $\widetilde{n}_j/\widetilde{n}_d$ are integers and $m\ge t_d=\widetilde{n}/\widetilde{n}_d^d$. We note that $t_d=1$ when $\widetilde{n}_j=\widetilde{n}_d$ for all j. Upper bounds for both $\ell_\pm(m)$ in the general setting are given in the proof of Theorem 2 in Subsection A3.3 of the Supplementary Material (Deng and Zhang (2020)).

To find upper bounds for $\ell_+(m)$, we partition $V_0=[\boldsymbol{a},\boldsymbol{b}]$ into an $\widetilde{n}_d\times\cdots\times\widetilde{n}_d$ lattice of small "unit blocks" of size $(\widetilde{n}_1/\widetilde{n}_d)\times\cdots\times(\widetilde{n}_d/\widetilde{n}_d)$, each composed of $t_d=\widetilde{n}/\widetilde{n}_d^d$ design points. Consider a line of such unit blocks $L_{\boldsymbol{k}}$ in the "anti-diagonal" direction and a region D_j between two contours of the unknown $f(\boldsymbol{x})$ at the levels c and $c+r_{q,+}^{1/q}(3^d m)$. In Figure 3, we color in red the unit blocks in $L_{\boldsymbol{k}}$ with nonempty intersection with D_j. Due to the monotonicity of the $\ell_+(m)$, it suffices to consider $m=k^d t_d$ for some integer $k\ge 1$. If $\boldsymbol{x}\le\boldsymbol{v}$ in $L_{\boldsymbol{k}}\cap D_j$ are separated by k unit blocks as depicted in Figure 3, then $m=k^d t_d<n_{\boldsymbol{x},\boldsymbol{v}}\le(k+2)^d t_d\le 3^d m$ and $f(\boldsymbol{v})-f(\boldsymbol{x})\le r_{q,+}^{1/q}(3^d m)\le r_{q,+}^{1/q}(n_{\boldsymbol{x},\boldsymbol{v}})$, so that $m_{\boldsymbol{x}}\ge n_{\boldsymbol{x},\boldsymbol{v}}>m$. Thus, the intersection contains no more than $(k+1)t_d\le 2m^{1/d}t_d^{1-1/d}$ design points $\boldsymbol{x}_i$ with $m_{\boldsymbol{x}_i}\le m$, all within k unit blocks from the upper contour. Let $J=\lceil\{f(\boldsymbol{b})-f(\boldsymbol{a})\}/r_{q,+}^{1/q}(3^d m)\rceil$. We divide $[\boldsymbol{a},\boldsymbol{b}]$ into J such regions D_j between consecutive contours with $\boldsymbol{a}\in D_1$ and $\boldsymbol{b}\in D_J$. The last region D_J is special. For $\boldsymbol{x}\in D_J$ with

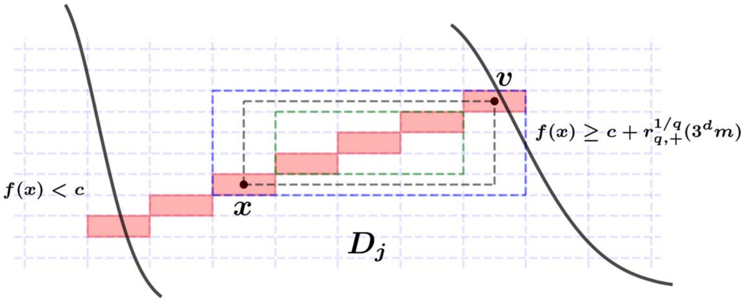


FIG. 3. *Upper bound for the number of design points with $m_{\boldsymbol{x}_i} \leq m$, an example: $d = 2$, $t_d = \widetilde{n}/\widetilde{n}_d^d$, $m = k^d t_d$ with $k = 3$, a line segment of unit blocks in the antidiagonal direction is colored in red, depicting its intersection with the region D_j between two contours of f; $\boldsymbol{x}$ is a design point k blocks away from the upper boundary of D_j, $\boldsymbol{v} \in D_j$; m, $n_{\boldsymbol{x},\boldsymbol{v}}$ and the upper bound $(k+2)^d m$ are respectively the number of points inside the rectangles colored in dashed green, gray and blue; as $m_{\boldsymbol{x}} \geq n_{\boldsymbol{x},\boldsymbol{v}} > m$ in this example, design points inside the intersection of D_j and these red unit blocks with $m_{\boldsymbol{x}_i} \leq m$ must belong to one of the $k+1 = 4$ upper-right unit blocks colored in red, and there are at most $(k+1)t_d = 4t_d$ such points in this example with $k = 3$. For general k and $m = k^d t_d$, $(k+1)t_d \leq 2m^{1/d} t_d^{1-1/d}$.*

$n_{\boldsymbol{x},\boldsymbol{b}} > m$, there must exist $\boldsymbol{v} \in [\boldsymbol{x}, \boldsymbol{b}]$ such that $m < n_{\boldsymbol{x},\boldsymbol{v}} \leq 2m$, so that $m_{\boldsymbol{x}} \geq n_{\boldsymbol{x},\boldsymbol{v}} > m$ due to $f(\boldsymbol{v}) \leq f(\boldsymbol{x}) + r_{q,+}^{1/q}(3^d m) \leq f(\boldsymbol{x}) + r_{q,+}^{1/q}(n_{\boldsymbol{x},\boldsymbol{v}})$. Thus, as there are no more than $d\widetilde{n}_d^{d-1}$ such $L_{\boldsymbol{k}}$ and $J - 1 \leq \{f(\boldsymbol{b}) - f(\boldsymbol{a})\}/r_{q,+}^{1/q}(3^d m) \leq \Delta_n^* \sigma / r_{q,+}^{1/q}(3^d m)$ regions D_j not containing $\boldsymbol{b}$, for $m = k^d t_d$ with integer $k \geq 1$,

$$\ell_+(m) \leq \min\left\{\widetilde{n}, d\widetilde{n}_d^{d-1}\left(\frac{\Delta_n^* \sigma}{r_{q,+}^{1/q}(3^d m)}\right)\left(2m^{\frac{1}{d}} t_d^{1-\frac{1}{d}}\right)\right\} + \#\{\boldsymbol{x}_i \in [\boldsymbol{a}, \boldsymbol{b}] : n_{\boldsymbol{x}_i,\boldsymbol{b}} \leq m\}$$

$$= \widetilde{n} \min\left\{1, m^{\frac{1}{d}+\frac{1}{2}}\left(\frac{\Delta_n^*}{\widetilde{n}^{1/d}}\right)\left(2d3^{d/2}/C_{q,d}^{1/q}\right)\right\} + \#\{\boldsymbol{x}_i \in [\boldsymbol{a}, \boldsymbol{b}] : n_{\boldsymbol{x}_i,\boldsymbol{b}} \leq m\}$$

with the variability bound $r_{q,+}(m) = C_{q,d}\sigma^q m^{-q/2}$ in (27). It follows that

$$\ell_\pm(m) \leq \ell_\pm^*(m) = \widetilde{n}\widetilde{H}(m) + \#\{\boldsymbol{x}_i \in [\boldsymbol{a}, \boldsymbol{b}] : n_{\boldsymbol{x}_i,\boldsymbol{b}} \leq m\} \quad \forall m \geq t_d \tag{40}$$

when $C_{q,d}^{1/q} \geq (2^{1/d+1/2})^d 2d3^{d/2}$. In Section A3.3 of the Supplementary Material (Deng and Zhang (2020)), we extend the above inequality to all $m \geq 1$ and prove (36) by applying (26) of Theorem 1 with the above $\ell_\pm^*(m)$ and the $r_{q,\pm}(m)$ in (27).

Theorem 2 is a comprehensive statement which gives rise to many conclusions. In the next sub-subsection, we prove that the block estimator is rate minimax in the ℓ_q risk for the entire lattice $[\boldsymbol{1}, \boldsymbol{n}]$ in a wide range of configurations of $\boldsymbol{n}$, q and Δ_n^*. In the next two subsections, we study the adaptation rate when $f(\cdot)$ is a piecewise constant function, and the variable selection rate when $f(\cdot)$ only depends on a subset of variables.

3.3.3. *Risk of the block estimator on the entire lattice and rate minimaxity.* We assume without loss of generality in this sub-subsection $n_1 \geq \cdots \geq n_d$. A direct comparison between Proposition 1 and Theorem 2 yields the following Theorem 3.

THEOREM 3. *Let $\widehat{f}_n^{(\text{block})}(\boldsymbol{x})$ be the block estimator in (16) with the lattice design $V = [\boldsymbol{1}, \boldsymbol{n}]$ as in (28). Assume ε_i are independent random variables with $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}|\varepsilon_i|^{q\vee 2} \leq \sigma^{q\vee 2}$. Let $s_q = \lceil 2/(q-1)\rceil \wedge (d+1)$, $n_s^* = \prod_{j=1}^s n_j$ for $s \leq d+1$ with $n_{d+1} = 1$, and*

$\Delta(\boldsymbol{f}_n/\sigma) = \{f(\boldsymbol{n}) - f(\mathbf{1})\}/\sigma$. *Then, for* $q \geq 1$,

$$\begin{aligned}&\sup\{R_q(\widehat{\boldsymbol{f}}_n^{(\mathrm{block})}, \boldsymbol{f}_n) : \boldsymbol{f}_n \in \mathcal{F}_n, \Delta(\boldsymbol{f}_n/\sigma) \leq \Delta_n^*\} \\ &\quad\lesssim_{q,d} \Lambda^{(\mathrm{match})} \inf_{\widehat{\boldsymbol{f}}} \sup_{\boldsymbol{f}_n \in \mathcal{F}_n} \{R_q(\widehat{\boldsymbol{f}}, \boldsymbol{f}_n) : \Delta(\boldsymbol{f}_n/\sigma) \leq \Delta_n^*\} + \frac{\sigma^q}{n}\left(\prod_{j=1}^d \log_+(n_j)\right)^{I\{q=2\}}\end{aligned} \tag{41}$$

holds when $\Delta_n^* \gtrsim_{q,d} t_{s_q}^{-1/2} = (n_{s_q}^*/n_{s_q}^{s_q})^{-1/2}$, *where* $\Lambda^{(\mathrm{match})} \leq \log n$ *is defined by*

$$\Lambda^{(\mathrm{match})} = \left[\log_+\left(\min\left\{\frac{n_{s_q}}{n_{s_q+1}}, \frac{n_{s_q}/(n_{s_q}^*)^{1/(s_q+2)}}{(\Delta_n^*)^{2/(s_q+2)}}\right\}\right)\right]^{I\{\frac{2}{q-1}=s_q\leq d, \Delta_n^*\leq n_{s_q}/t_{s_q}^{1/2}\}}. \tag{42}$$

Moreover, when $\max_{j\leq d} n_j \lesssim_d n^{1/d}$ *and* $\Delta(\boldsymbol{f}_n/\sigma) \leq \Delta_n^*$,

$$\begin{aligned}&R_q(\widehat{\boldsymbol{f}}_n^{(\mathrm{block})}, \boldsymbol{f}_n) \\ &\quad\lesssim_{q,d} \sigma^q \min\left\{1, \left(\frac{\Delta_n^*}{n^{1/d}}\right)^{\min\{1, \frac{qd}{d+2}\}}\left[\log_+\left(n \wedge \left(\frac{n^{1/d}}{\Delta_n^*}\right)^{2d/(d+2)}\right)\right]^{\delta_1} + \frac{(\log n)^{d\delta_2}}{n^{(q/2)\wedge 1}}\right\},\end{aligned} \tag{43}$$

holds for all $\Delta_n^* \geq 0$, *where* $\delta_1 = I\{\frac{qd}{d+2} = 1\}$ *and* $\delta_2 = I\{q = 2\}$.

REMARK 4. It can be seen in our analysis that the logarithmic term presents for $q = 2$, as the last component on the right-hand side of (36), (41) and (43), due to the lack of data near the extreme points $\{\boldsymbol{a}, \boldsymbol{b}\}$ or $\{\mathbf{1}, \boldsymbol{n}\}$ of the domain.

Compared with Proposition 1, Theorem 3 shows that the risk of the block estimator matches the minimax rate when $\Delta_n^* \geq t_{s_q}^{-1/2} = (\prod_{j=1}^{s_q}(n_j/n_{s_q}))^{-1/2}$ ($\Delta_n^* \geq n^{-1/2}$ if $s_q = d + 1$) possibly up to a logarithmic factor $\Lambda^{(\mathrm{match})} \leq \log(n)$, provided that the minimax rate is no faster than $\sigma^q n^{-1}(\prod_{j=1}^d \log_+(n_j))^{\delta_2}$ due to the edge effect. The match is always exact when $2/(q - 1) \neq s_q \leq d$, that is, $2/(q - 1)$ is not an integer or an integer greater than d. When $2/(q - 1) = s_q \leq d - 1$ and $n_{s_q} \asymp n_{s_q+1}$, $\Lambda^{(\mathrm{match})} = O(1)$ and the match is also exact. However, in the interesting setting where $q = d = 2$ and $n_1 \asymp n_2$, we have $s_q = 2$ so that $\Lambda^{(\mathrm{match})} \asymp \log(n)$ when $\Delta_n^* \ll n_2$.

The one-dimensional risk bound for all $q \geq 1$ can be obtained from (43) as

$$T_q([1, n]) \lesssim_q \sigma^q n \min\left\{1, \left(\frac{\Delta_n^*}{n}\right)^{\min\{\frac{q}{3}, 1\}}\left[\log_+\left(n \wedge \left(\frac{n}{\Delta_n^*}\right)^{\frac{2}{3}}\right)\right]^{I\{q=3\}} + \frac{(\log_+(n))^{I\{q=2\}}}{n^{(q/2)\wedge 1}}\right\},$$

which reproduces (4) for $1 \leq q < 3$. We note that if we view one-dimensional isotonic regression as multidimensional on an $n_1 \times 1 \times \cdots \times 1$ lattice, the general bound yields this one-dimensional $n_1^{-1/3}$-rate. Interestingly, for general $\boldsymbol{n}$, we still have the one-dimensional rate as long as the effective dimension s is 0 or 1, that is, $\Delta_n^* \geq n_2/t_2^{1/2} = n_2^{3/2}/n_1^{1/2}$. For $q = 2$ and $d \geq 2$, it follows from Theorem 2 that when $\Delta_n^* \geq n_2/t_2^{1/2} = n_2^{3/2}/n_1^{1/2}$, we have $s < s_q = 2$ and only the first two cases of (37) are effective. This implies

$$T_2([\mathbf{1}, \boldsymbol{n}]) \lesssim_d \sigma^2 n \min\left\{1, (\Delta_n^*/n_1)^{2/3} + \prod_{j=1}^d (\log_+(n_j)/n_j)\right\},$$

exactly the same as the bound of $T_2([1, n_1])$ in univariate case when $(\Delta_n^*/n_1)^{2/3}$ is dominant in both rates. In this case, our theory does not guarantee an advantage of the multiple isotonic regression on the entire lattice in terms of the ℓ_2 risk, compared with the row-by-row univariate isotonic regression of length n_1. This observation agrees with Chatterjee, Guntuboyina and Sen (2018) where the ℓ_2 minimax rate of two-dimensional isotonic regression, $\sigma^2\Delta_n^* n^{-1/2}$, requires $n_2^{3/2}/n_1^{1/2} \geq \Delta_n^*$.

To conclude this subsection, we compare the ℓ_2 risk bound for the block estimator in Theorem 3 with those for the LSE in the existing literature. For $d = 2$, Chatterjee, Guntuboyina and Sen (2018) gives an upper bound for the LSE as

$$R_2(\widehat{\boldsymbol{f}}_n^{(\mathrm{lse})}, \boldsymbol{f}_n) \lesssim \sigma^2\biggl(\frac{\Delta_n^*}{\sqrt{n}}(\log n)^4 + \frac{1}{n}(\log n)^8\biggr)$$

for any $n_1 \times n_2$ lattice and f satisfying $\Delta(\boldsymbol{f}_n/\sigma) \leq n_2^{3/2}/n_1^{1/2}$, in contrast to

$$R_2(\widehat{\boldsymbol{f}}_n^{(\mathrm{block})}, \boldsymbol{f}_n) \lesssim \sigma^2\biggl(\frac{\Delta_n^*}{\sqrt{n}}\log(n) + \frac{1}{n}(\log n)^2\biggr)$$

in (43) of Theorem 3 or in the third case of (37) of Theorem 2 with $[\boldsymbol{a}, \boldsymbol{b}] = [\mathbf{1}, \boldsymbol{n}]$. However, for $n_1 = \cdots = n_d = n^{1/d}$ and $\Delta_n^* = 1$ as in Han et al. (2019) for $d \geq 3$, (43) is reduced to

$$R_2(\widehat{\boldsymbol{f}}_n^{(\mathrm{block})}, \boldsymbol{f}_n) \lesssim_d n^{-1/d},$$

which should be compared with the the rate

$$R_2(\widehat{\boldsymbol{f}}_n^{(\mathrm{lse})}, \boldsymbol{f}_n) \lesssim_d n^{-1/d}\log^4(n)$$

for the LSE (Han et al. (2019)).

3.4. *Adaptation rate of the block estimator with lattice designs in the piecewise constant case.* We consider here the adaptation behavior of the block estimator in the setting where $f(\cdot)$ is piecewise constant on a union of rectangles, as a direct consequence of Theorem 2.

THEOREM 4. *Let $\widehat{f}_n^{(\mathrm{block})}(\boldsymbol{x})$ be the block estimator in* (16). *Assume ε_i are independent variables with $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}|\varepsilon_i|^{q\vee 2} \leq \sigma^{q\vee 2}$ and f is nondecreasing and piecewise constant on V in the sense of $V = \bigcup_{k=1}^K [\boldsymbol{a}_k, \boldsymbol{b}_k]$ with $K \leq n$ and $f(\boldsymbol{a}_k) = f(\boldsymbol{b}_k)$ for all $k \leq K$. Then*

$$R_q(\widehat{\boldsymbol{f}}_n^{(\mathrm{block})}, \boldsymbol{f}_n) \lesssim_{q,d} \sigma^q \min\Biggl\{1, n^{-1}\sum_{k=1}^K n_{\boldsymbol{a}_k,\boldsymbol{b}_k}^{(1-q/2)_+}\bigl(\log_+^{s_k}(n_{\boldsymbol{a}_k,\boldsymbol{b}_k})\bigr)^{I\{q=2\}}\Biggr\}$$

with $s_k = \#\{j : b_{k,j} > a_{k,j}\}$. Moreover, if in addition $\{[\boldsymbol{a}_k, \boldsymbol{b}_k], k = 1, \ldots, K\}$ are disjoint, then

$$R_q(\widehat{\boldsymbol{f}}_n^{(\mathrm{block})}, \boldsymbol{f}_n) \lesssim_{q,d} \sigma^q \min\biggl\{1, \Bigl(\frac{K}{n}\Bigr)^{\min\{1,q/2\}}\bigl(\log_+^{d_K}(n/K)\bigr)^{I\{q=2\}}\biggr\}, \tag{44}$$

where $d_K = \max_{1\leq k\leq K} s_k$ is the largest dimension of $[\boldsymbol{a}_k, \boldsymbol{b}_k]$ in the partition.

The rate in (44) is consistent with existing results for $d = 1$ under which the block estimator is the LSE and the mean squared risk bound is

$$R_2(\widehat{\boldsymbol{f}}_n^{(\mathrm{block})}, \boldsymbol{f}_n) \lesssim \sigma^2\frac{K}{n}\log_+(n/K).$$

In general, the risk bound in (44) under $q = 2$ is reduced to at most

$$\sigma^2\frac{K}{n}\log_+^d(n/K),$$

which should be compared with

$$\sigma^2\left(\frac{K}{n}\right)^{2/d}\log_+^8(n/K)$$

for the LSE as in Chatterjee, Guntuboyina and Sen (2018) for $d=2$ and in Han et al. (2019) for $d\geq 3$.

REMARK 5. Han et al. (2019) proved that even when $f(\cdot)$ is a constant function, that is, $K=1$,

$$R_2(\widehat{\boldsymbol{f}}_n^{(\mathrm{lse})},\boldsymbol{f}_n)\gtrsim_d \sigma^2 n^{-2/d}$$

so the adaptation rate of the LSE, $(K/n)^{2/d}$, cannot be further improved, which means the LSE is unable to adapt to parametric rate for $d\geq 3$.

The adaptation rate in (44) also implies that when $[\boldsymbol{a}_k,\boldsymbol{b}_k]$ are two-dimensional sheets (i.e., $|\{j:b_{k,j}\neq a_{k,j}\}|\leq 2$), the upper bound turns out to be

$$\frac{K}{n}\log_+^2(n/K),$$

which again should be compared with

$$\frac{K}{n}\log_+^8(n/K)$$

in Han et al. (2019).

3.5. *Adaptive estimation to variable selection with lattice designs.* In this subsection, we consider the case where the true function of interest, $f(\cdot)$, depends only on a subset S of s variables, that is, $f(\boldsymbol{x})=f_S(\boldsymbol{x}_S)$. We study the adaptive estimation when $\max_{j\leq d} n_j\lesssim_d n^{1/d}$, that is, $n_j\asymp n^{1/d}$ for all $1\leq j\leq d$.

THEOREM 5. *Assume $f(\cdot)$ is nondecreasing and dependent only on an unknown set S of $s<d$ variables. Let $\widehat{f}_n^{(\mathrm{block})}(\boldsymbol{x})$ be the block estimator in (16) on the lattice design $V=[\mathbf{1},\boldsymbol{n}]$. Assume $\max_{1\leq j\leq d} n_j\lesssim_d n^{1/d}$ and ε_i's are independent and satisfies $\mathbb{E}\varepsilon_i=0$ and $\mathbb{E}|\varepsilon_i|^{q\vee 2}\leq\sigma^{q\vee 2}$. Let $\Delta(\boldsymbol{f}_n/\sigma)=\{f(\boldsymbol{n})-f(\mathbf{1})\}/\sigma$. Then*

$$\begin{aligned}&\sup\{R_q(\widehat{\boldsymbol{f}}_n^{(\mathrm{block})},\boldsymbol{f}_n):\boldsymbol{f}_n\in\mathcal{F}_n, f(\boldsymbol{x})=f_S(\boldsymbol{x}_S),\Delta(\boldsymbol{f}_n/\sigma_S)\leq\Delta^*_{n,S}\}\\ &\quad\lesssim_d \sigma_S^q\min\{\Lambda_{s,1}^{(\mathrm{select})},\Lambda_{s,2}^{(\mathrm{select})}(\Delta^*_{n,S}/n^{1/d})^{\min\{1,\frac{qs}{s+2}\}}\\ &\qquad+\Lambda_{s,1}^{(\mathrm{select})}(n^{s/d})^{-\min\{1,q/2\}}(\log n)^{sI\{q=2\}}\}\end{aligned}\tag{45}$$

for all $1\leq s\leq d$, where $\sigma_S=\sigma/(\prod_{j\notin S}n_j)^{1/2}\leq C_d\sigma/n^{(1-s/d)/2}$ and

$$\Lambda_{s,1}^{(\mathrm{select})}=\left(\sum_{j=1}^{n^{1/d}}j^{-q/2}/(n^{1/d})^{1-q/2}\right)^{d-s},$$

$$\Lambda_{s,2}^{(\mathrm{select})}=\left(\sum_{j=1}^{n^{1/d}}j^{\min\{\frac{1-q}{2},-\frac{q}{s+2}\}}/(n^{1/d})^{\min\{\frac{1-q}{2},-\frac{q}{s+2}\}+1}\right)^{d-s}(\log n)^{I\{\frac{qs}{s+2}=1\}}.$$

In particular,

$$R_2(\widehat{\boldsymbol{f}}_n^{(\text{block})}, \boldsymbol{f}_n) \lesssim_d \begin{cases} \sigma^2 n^{s/d-1} \min\{(\log n)^{d-s}, \Delta^*_{n,S} n^{-1/d} (\log n)^{I\{s=2\}} + n^{-s/d} (\log n)^d\}, & s \geq 2, \\ \sigma^2 n^{s/d-1} \min\{(\log n)^{d-1}, (\Delta^*_{n,S}/n^{1/d})^{2/3} + n^{-1/d} (\log n)^d\}, & s = 1. \end{cases} \tag{46}$$

In the proof of Theorem 5, the key observation is that in the sheet of $\boldsymbol{x}$ with fixed $\boldsymbol{x}_{S^c}$, the risk bound is identical to that of model S with σ^q reduced by a factor of $n^{-q/2}_{\boldsymbol{x}_{S^c}, \boldsymbol{n}_{S^c}}$. The above rate would then become clear after the summation of risk bounds over $\boldsymbol{x}_{S^c}$.

Let $n_j = n^{1/d}$ for all j. Consider an oracle expert with the extra knowledge of the subset S. Suppose the oracle expert first computes the average of the $n^{1-s/d}$ values of y_i holding $\boldsymbol{x}_S$ fixed and then solves the s-dimensional isotonic regression problem at the noise level $\sigma_S = \sigma n^{(s/d-1)/2}$. For this oracle expert, the sample size becomes $n^{s/d}$ and the condition on the range-to-noise ratio becomes $(f(\boldsymbol{n}) - f(\mathbf{1}))/\sigma_S \leq \Delta^*_{n,S}$, equivalent to $(f(\boldsymbol{n}) - f(\mathbf{1}))/\sigma \leq \Delta^*_n$ with $\Delta^*_{n,S} = \Delta^*_n n^{(1-s/d)/2}$. It follows from (30) in Proposition 1 that for $\varepsilon_i \sim N(0, \sigma^2)$ and $\Delta^*_{n,S} \geq (n^{-(s/d)/2}) \vee (I\{q > 1 + 2/s\})$, the ℓ_q minimax lower bound for the oracle expert is

$$\begin{aligned} &\inf_{\widehat{\boldsymbol{f}}} \sup\{R_q(\widehat{\boldsymbol{f}}, \boldsymbol{f}_n) : \boldsymbol{f}_n \in \mathcal{F}_n, f(\boldsymbol{x}) = f_S(\boldsymbol{x}_S), \Delta(\boldsymbol{f}_n/\sigma_S) \leq \Delta^*_{n,S}\} \\ &\quad \gtrsim \sigma_S^q \min\{1, (\Delta^*_{n,S}/n^{1/d})^{\min\{1, qs/(s+2)\}}\}. \end{aligned}$$

Hence the variable-selection adaptation rate in (45) matches the oracle minimax lower bound up to some constant or logarithmic factors $\Lambda^{(\text{select})}_{s,1}$, $\Lambda^{(\text{select})}_{s,2}$ and $\Lambda^{(\text{select})}_{s,1} (\log n)^{sI\{q=2\}}$, provided that

$$\Delta^*_{n,S} \geq \max(n^{-s/(2d)}, I\{q > 1 + 2/s\}),$$

or equivalently $\Delta(\boldsymbol{f}_n/\sigma) \leq \Delta^*_n$ with $\Delta^*_n \geq \max(n^{-1/2}, n^{-(1-s/d)/2} I\{q > 1 + 2/s\})$. The match to the oracle minimax rate is always exact for $q = 1$ and any s as both $\Lambda^{(\text{select})}_{s,1}$ and $\Lambda^{(\text{select})}_{s,2}$ are bounded by a constant. When $q = 2$, the match is also exact but up to some logarithmic factors as $\Lambda^{(\text{select})}_{s,1} \lesssim_d (\log n)^{d-s}$ and $\Lambda^{(\text{select})}_{s,2} \lesssim_d (\log n)^{I\{s=2\}}$.

3.6. *Multiple isotonic regression with random designs.* In this subsection, we consider $V = [\mathbf{0}, \mathbf{1}]$ in continuum and, same as before, $\boldsymbol{a} \preceq \boldsymbol{b}$ iff $\boldsymbol{a} \leq \boldsymbol{b}$. Different from fixed designs, here $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are i.i.d. random vectors from a distribution $\mathbb{P}$ supported on $[\mathbf{0}, \mathbf{1}]$. For simplicity, we assume the distribution of the design points has a Lebesgue density bounded both from above and below; for $\mu_{\boldsymbol{u},\boldsymbol{v}} = \mathbb{P}\{\boldsymbol{u} \leq \boldsymbol{x}_i \leq \boldsymbol{v}\}$ and the Lebesgue $\mu^L_{\boldsymbol{u},\boldsymbol{v}} = \mu^L([\boldsymbol{u}, \boldsymbol{v}]) = \int_{[\boldsymbol{u},\boldsymbol{v}]} d\boldsymbol{x}$,

$$\rho_1 \mu^L_{\boldsymbol{u},\boldsymbol{v}} \leq \mu_{\boldsymbol{u},\boldsymbol{v}} \leq \rho_2 \mu^L_{\boldsymbol{u},\boldsymbol{v}}, \tag{47}$$

with certain fixed constants $0 < \rho_1 \leq \rho_2 < \infty$. We consider the integrated L_q risk in (3), that is,

$$R^*_q(\widehat{f}_n^{(\text{block})}, f) = \int_{\boldsymbol{x} \in [\mathbf{0},\mathbf{1}]} \mathbb{E}|\widehat{f}_n^{(\text{block})}(\boldsymbol{x}) - f(\boldsymbol{x})|^q \, d\boldsymbol{x},$$

and partial integrated L_q risk on block $[\boldsymbol{a}, \boldsymbol{b}]$ as

$$R^*_q([\boldsymbol{a}, \boldsymbol{b}]) = \int_{[\boldsymbol{a},\boldsymbol{b}]} \mathbb{E}|\widehat{f}_n^{(\text{block})}(\boldsymbol{x}) - f(\boldsymbol{x})|^q \, d\boldsymbol{x}.$$

THEOREM 6. *Let $\widehat{f}_n^{(\text{block})}(\boldsymbol{x})$ be the block estimator in (16) with $V = [\boldsymbol{0}, \boldsymbol{1}]$. Assume $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in [\boldsymbol{0}, \boldsymbol{1}]$ are i.i.d. random vectors drawn from a distribution satisfying (47). Assume f is nondecreasing and ε_i are independent random variables with $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}|\varepsilon_i|^{q\vee 2} \le \sigma^{q\vee 2}$. Let $\{\boldsymbol{a}, \boldsymbol{b}\} \subset V$ with $\boldsymbol{a} \le \boldsymbol{b}$. Then, for $q \ge 1$,*

$$
\begin{aligned}
R_q^*([\boldsymbol{a}, \boldsymbol{b}]) &= \int_{[\boldsymbol{a}, \boldsymbol{b}]} \mathbb{E}|\widehat{f}_n^{(\text{block})}(\boldsymbol{x}) - f(\boldsymbol{x})|^q \, d\boldsymbol{x} \\
&\le C^*_{q,d,\rho_1,\rho_2} \sigma^q \biggl[\int_0^{n\mu_{\boldsymbol{a},\boldsymbol{b}}} \bigl((t \vee 1)^{-q/2} + \Delta^q_{\boldsymbol{a},\boldsymbol{b}} e^{-t}\bigr) H^*(dt) \qquad (48) \\
&\qquad + \int_{\boldsymbol{x} \in [\boldsymbol{a}, \boldsymbol{b}]} \bigl(\{(n\mu_{\boldsymbol{x},\boldsymbol{b}}) \vee 1\}^{-q/2} + \Delta^q_{\boldsymbol{0},\boldsymbol{1}} e^{-n\mu_{\boldsymbol{x},\boldsymbol{b}}}\bigr) \, d\boldsymbol{x} \biggr],
\end{aligned}
$$

where $\Delta_{\boldsymbol{u},\boldsymbol{v}} = (f(\boldsymbol{v}) - f(\boldsymbol{u}))/\sigma$ and $\mu_{\boldsymbol{u},\boldsymbol{v}} = \mathbb{P}\{\boldsymbol{x}_i \in [\boldsymbol{u}, \boldsymbol{v}]\}$ for all $\boldsymbol{u} \le \boldsymbol{v}$ and $H^*(t) = \min\{1, \Delta_{\boldsymbol{a},\boldsymbol{b}} (n\mu_{\boldsymbol{a},\boldsymbol{b}})^{-1/d} t^{1/2+1/d}\}$. Specifically, (48) is no greater than

$$
\begin{aligned}
&\sigma^q \min\biggl\{ (\Delta^q_{\boldsymbol{0},\boldsymbol{1}} + 1)\mu_{\boldsymbol{a},\boldsymbol{b}}, \biggl(\frac{\Delta_{\boldsymbol{a},\boldsymbol{b}}}{(n\mu_{\boldsymbol{a},\boldsymbol{b}})^{1/d}} \biggr)^{\min\{1, \frac{qd}{d+2}\}} \Lambda_1^{(\text{random})} \qquad (49) \\
&\qquad + \frac{\Delta^{q+1}_{\boldsymbol{a},\boldsymbol{b}}}{(n\mu_{\boldsymbol{a},\boldsymbol{b}})^{1/d}} + (\Delta^q_{\boldsymbol{0},\boldsymbol{1}} + 1)\mu_{\boldsymbol{a},\boldsymbol{b}} \frac{\Lambda_2^{(\text{random})}}{(n\mu_{\boldsymbol{a},\boldsymbol{b}})^{(q/2)\wedge 1}} \biggr\}
\end{aligned}
$$

up to a constant depending on q, d, ρ_1, ρ_2 only, where

$$
\Lambda_1^{(\text{random})} = \bigl[\log_+\bigl(n\mu_{\boldsymbol{a},\boldsymbol{b}} \wedge \bigl((n\mu_{\boldsymbol{a},\boldsymbol{b}})^{\frac{2}{d+2}} / \Delta_{\boldsymbol{a},\boldsymbol{b}}^{2d/(d+2)}\bigr)\bigr)\bigr]^{I\{\frac{qd}{d+2}=1\}}
$$

and $\Lambda_2^{(\text{random})} = (\log_+(n\mu_{\boldsymbol{a},\boldsymbol{b}}))^{dI\{q=2\}+(d-1)I\{q>2\}}$.

The $H^*(t)$ here is identical to the $\widetilde{H}(t)$ in Theorem 2 in $t \in [t_d, n]$, effectively taking $t_d = 1$. This reveals an intrinsic difference between lattice design and random design: the effective dimension of the random design over $[\boldsymbol{a}, \boldsymbol{b}] \subseteq [\boldsymbol{0}, \boldsymbol{1}]$ is always d—any hyperrectangle $[\boldsymbol{a}, \boldsymbol{b}]$ with positive measure behaves similar to a hypercube. The above rate in (49) is therefore comparable to the rate in (43) for the lattice design with $n_j = n^{1/d}$ for all j. In fact, the rate in (49) can be derived from a scale change of the upper bound for $R_q^*([\boldsymbol{0}, \boldsymbol{1}])$.

The study of the integrated L_q risk in isotonic regression is relatively new. Fokianos, Leucht and Neumann (2017) gives an asymptotic bound, $O(n^{-1/(d+2)})$, for the L_1 risk with $[\boldsymbol{a}, \boldsymbol{b}] = [\boldsymbol{0}, \boldsymbol{1}]$. The L_1 error bound in Theorem 6 is consistent with their result.

To fit in with random design, we now define $r_{q,+}(m)$ as a nonincreasing function of $m \in [0, n]$ in continuum satisfying

$$
r_{q,+}(m) \ge \max\biggl\{ \mathbb{E}\biggl(\max_{\boldsymbol{u} \preceq \boldsymbol{x}} \sum_{\boldsymbol{x}_i \in [\boldsymbol{u}, \boldsymbol{v}]} \frac{\varepsilon_i}{n_{\boldsymbol{u},\boldsymbol{v}} \vee 1} \biggr)_+^q : \mathbb{E}[n_{\boldsymbol{x},\boldsymbol{v}}] = m, \boldsymbol{x} \preceq \boldsymbol{v} \text{ and } \boldsymbol{v} \in V_0 \biggr\}, \qquad (50)
$$

and modify the definition of $m_{\boldsymbol{x}} = m_{\boldsymbol{x},+}$ in (23) to

$$
m_{\boldsymbol{x}} = n\mu_{\boldsymbol{x},\boldsymbol{v}_{\boldsymbol{x}}} \quad \text{where } \boldsymbol{v}_{\boldsymbol{x}} = \operatorname*{arg\,sup}_{\boldsymbol{x} \le \boldsymbol{v} \le \boldsymbol{b}} \{n\mu_{\boldsymbol{x},\boldsymbol{v}} : f(\boldsymbol{v}) \le f(\boldsymbol{x}) + r_{q,+}^{1/q}(n\mu_{\boldsymbol{x},\boldsymbol{v}})\}. \qquad (51)
$$

Note $n_{\boldsymbol{x},\boldsymbol{v}}$, the number of design points in $[\boldsymbol{x}, \boldsymbol{v}]$, becomes a Binomial$(n, \mu_{\boldsymbol{x},\boldsymbol{v}})$ random variable. Here, we omit $m_{\boldsymbol{x},-}$ as it can be analyzed by symmetry. Nevertheless, Theorem 6 is still proved in a similar way to Theorem 2. However, different from (25) in Theorem 1, the point risk bound is given by the following proposition.

PROPOSITION 3. *Assume the conditions of Theorem 6. Then (50) holds for*

$$r_{q,+}(n\mu_{\boldsymbol{x},\boldsymbol{v}}) = C_{q,d,\rho_1,\rho_2}\sigma^q(n\mu_{\boldsymbol{x},\boldsymbol{v}} \vee 1)^{-q/2} \tag{52}$$

with C_{q,d,ρ_1,ρ_2} continuous in $q \in [1,\infty)$ and for all $\boldsymbol{x} \in [\boldsymbol{a},\boldsymbol{b}]$,

$$\begin{aligned}&\mathbb{E}(\widehat{f}_n^{(\text{block})}(\boldsymbol{x}) - f(\boldsymbol{x}))_+^q\\ &\quad\le 2^q r_{q,+}(m_{\boldsymbol{x}}) + 2^{q-1}\sigma^q C_{q,d,\rho_1,\rho_2}((\Delta_{\boldsymbol{a},\boldsymbol{b}}^q + 1)e^{-m_{\boldsymbol{x}}} + (\Delta_{\boldsymbol{0},\boldsymbol{1}}^q + 1)e^{-n\mu_{\boldsymbol{x},\boldsymbol{b}}}).\end{aligned} \tag{53}$$

As we discussed below (23), the positive part of the bias of $\widehat{f}_n^{(\text{block})}(\boldsymbol{x})$ is of no greater order than the variability of the noise as measured by $r_{q,+}^{1/q}(n_{\boldsymbol{x},\boldsymbol{v}_{\boldsymbol{x}}}) \asymp r_{q,+}^{1/q}(m_{\boldsymbol{x}})$ provided the presence of at least one design point in $[\boldsymbol{x},\boldsymbol{v}_{\boldsymbol{x}}]$. The first term on the right-hand side of (53) thus comes from the case of $n_{\boldsymbol{x},\boldsymbol{v}_{\boldsymbol{x}}} > 0$. However, $[\boldsymbol{x},\boldsymbol{v}_{\boldsymbol{x}}]$ might be an empty cell with no design points. We then have to consider points in $[\boldsymbol{x},\boldsymbol{b}]$ when $n_{\boldsymbol{x},\boldsymbol{v}_{\boldsymbol{x}}} = 0$ and in $[\boldsymbol{x},\boldsymbol{1}]$ when $n_{\boldsymbol{x},\boldsymbol{b}} = 0$, leading to terms with $\Delta_{\boldsymbol{a},\boldsymbol{b}}$ and $\Delta_{\boldsymbol{0},\boldsymbol{1}}$, respectively.

Corresponding to Theorems 3 and 4, the following two theorems give the risk bounds for random designs under the general case and the piecewise constant case for the entire $[\boldsymbol{0},\boldsymbol{1}]$. Due to space limitations, the minimax rate and the adaptation rate to variable selection in random design are not discussed.

THEOREM 7. *Let $\widehat{f}_n^{(\text{block})}(\boldsymbol{x})$, f and $\{\boldsymbol{x}_i,\varepsilon_i, i \le n\}$ be as in Theorem 6. Suppose $\Delta_{\boldsymbol{0},\boldsymbol{1}} = (f(\boldsymbol{1}) - f(\boldsymbol{0}))/\sigma$ is bounded by a constant. Then*

$$\begin{aligned}R_q^*(\widehat{f}_n^{(\text{block})}, f) \lesssim_{q,d,\rho_1,\rho_2} &\ \sigma^q\left(\frac{\Delta_{\boldsymbol{0},\boldsymbol{1}}}{n^{1/d}}\right)^{\min\{1,\frac{qd}{d+2}\}}(\log n)^{I\{\frac{qd}{d+2}=1\}}\\ &+ \frac{(\log n)^{dI\{q=2\}+(d-1)I\{q>2\}}}{n^{(q/2)\wedge 1}}.\end{aligned}$$

In particular, when $q = 2$ and $d \ge 2$,

$$R_2^*(\widehat{f}_n^{(\text{block})}, f) \lesssim_{d,\rho_1,\rho_2} \sigma^2 \min\left\{1, \frac{\Delta_{\boldsymbol{0},\boldsymbol{1}}}{n^{1/d}}(\log n)^{I\{d=2\}} + \frac{(\log n)^d}{n}\right\}. \tag{54}$$

REMARK 6. For simplicity, we here consider the case of bounded $\Delta_{\boldsymbol{0},\boldsymbol{1}}$. Theorem 6 also directly yields error bounds for general $\Delta_{\boldsymbol{0},\boldsymbol{1}}$ by setting $[\boldsymbol{a},\boldsymbol{b}] = [\boldsymbol{0},\boldsymbol{1}]$ in (48) and (49).

THEOREM 8. *Let $\widehat{f}_n^{(\text{block})}(\boldsymbol{x})$, f and $\{\boldsymbol{x}_i,\varepsilon_i, i \le n\}$ be as in Theorem 6. Suppose V has disjoint partition $V = \bigcup_{k=1}^K [\boldsymbol{a}_k,\boldsymbol{b}_k]$ with $K \le n$ and $f(\boldsymbol{a}_k) = f(\boldsymbol{b}_k)$ for all $k \le K$. Then*

$$\begin{aligned}&R_q^*(\widehat{f}_n^{(\text{block})}, f)\\ &\quad\lesssim_{q,d,\rho_1,\rho_2} \sigma^q(\Delta_{\boldsymbol{0},\boldsymbol{1}}^q + 1)\left(\frac{K}{n}\right)^{\min\{1,q/2\}}(\log_+(n/K))^{dI\{q=2\}+(d-1)I\{q>2\}},\end{aligned} \tag{55}$$

where $\Delta_{\boldsymbol{0},\boldsymbol{1}} = (f(\boldsymbol{1}) - f(\boldsymbol{0}))/\sigma$. In particular, when $q = 2$,

$$R_2^*(\widehat{f}_n^{(\text{block})}, f) \lesssim_{d,\rho_1,\rho_2} \sigma^2(\Delta_{\boldsymbol{0},\boldsymbol{1}}^2 + 1)\frac{K}{n}\log_+^d(n/K).$$

We can also derive risk bounds for the empirical ℓ_q risk. As $[\boldsymbol{x}_i,\boldsymbol{v}_{\boldsymbol{x}_i}]$ always has the design point $\boldsymbol{x}_i$, there is no "empty cell" problem as in Proposition 3 when bounding the empirical risk. It follows that

$$\mathbb{E}[(\widehat{f}_n^{(\text{block})}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i))_+^q|\boldsymbol{x}_i = \boldsymbol{x}] \lesssim_{q,d,\rho_1,\rho_2} r_{q,+}(m_{\boldsymbol{x}}),$$

so that

$$R_q(\widehat{\boldsymbol{f}}_n^{(\text{block})}, \boldsymbol{f}) \\ \lesssim_{q,d,\rho_1,\rho_2} \sigma^q \min\left\{\mu_{\boldsymbol{a},\boldsymbol{b}}, \left(\frac{\Delta_{\boldsymbol{a},\boldsymbol{b}}}{(n\mu_{\boldsymbol{a},\boldsymbol{b}})^{1/d}}\right)^{\min\{1,\frac{qd}{d+2}\}} \Lambda_1^{(\text{random})} + \mu_{\boldsymbol{a},\boldsymbol{b}} \frac{\Lambda_1^{(\text{random})}}{(n\mu_{\boldsymbol{a},\boldsymbol{b}})^{(q/2)\wedge 1}}\right\}$$

by an almost identical proof. It follows that under the conditions of Theorem 6 and $\Delta_{\mathbf{0},\mathbf{1}} = 1$, the worst case upper bound of the mean squared risk is

$$R_2(\widehat{\boldsymbol{f}}_n^{(\text{block})}, \boldsymbol{f}) \lesssim_{d,\rho_1,\rho_2} \sigma^2 n^{-1/d} (\log n)^{I\{d=2\}},$$

and under the conditions of Theorem 8, the mean squared risk bound in piecewise constant case is

$$R_2(\widehat{\boldsymbol{f}}_n^{(\text{block})}, \boldsymbol{f}) \lesssim_{d,\rho_1,\rho_2} \sigma^2 \frac{K}{n} \log^d(n/K).$$

We shall compare the above two rates with the results for the LSE in Han et al. (2019) respectively, that is,

$$\sigma^2 n^{-1/d} \log^{\gamma_d}(n)$$

and

$$\sigma^2 \left(\frac{K}{n}\right)^{2/d} \log^{2\gamma_d}(en/K),$$

where $\gamma_2 = 9/2$ and $\gamma_d = (d^2 + d + 1)/2$ when $d \geq 3$. It is worth mentioning that Han et al. (2019) also proved the piecewise constant rate for the LSE, $(K/n)^{2/d}$, is not improvable as when $K = 1$,

$$R_2(\widehat{\boldsymbol{f}}_n^{(\text{lse})}(\boldsymbol{x}), \boldsymbol{f}) \gtrsim_{d,\rho_1,\rho_2} \sigma^2 n^{-2/d}.$$

3.7. *Model misspecification.* We consider in this subsection properties of the block estimator in the nonparametric regression model

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{56}$$

for general f. When the true regression function f fails to be nondecreasing, the isotonic regression model (1) is misspecified, so that the block estimators actually estimate their noiseless versions, say $\overline{f}_n^*(\boldsymbol{x})$, instead of the true f. For the block max-min and min-max estimator in (10),

$$\begin{aligned} \overline{f}_n^*(\boldsymbol{x}) = \overline{f}_n^{(\text{max-min})}(\boldsymbol{x}) &= \max_{\boldsymbol{u} \preceq \boldsymbol{x}, n_{\boldsymbol{u},*} > 0} \min_{\boldsymbol{x} \preceq \boldsymbol{v}, n_{\boldsymbol{u},\boldsymbol{v}} > 0} \overline{f}_{[\boldsymbol{u},\boldsymbol{v}]} \quad \forall \boldsymbol{x} \in V, \\ \overline{f}_n^*(\boldsymbol{x}) = \overline{f}_n^{(\text{min-max})}(\boldsymbol{x}) &= \min_{\boldsymbol{x} \preceq \boldsymbol{v}, n_{*,\boldsymbol{v}} > 0} \max_{\boldsymbol{u} \preceq \boldsymbol{x}, n_{\boldsymbol{u},\boldsymbol{v}} > 0} \overline{f}_{[\boldsymbol{u},\boldsymbol{v}]} \quad \forall \boldsymbol{x} \in V, \end{aligned} \tag{57}$$

are their noiseless versions, where $\overline{f}_A$ denotes the average of $\{f(\boldsymbol{x}_i) : 1 \leq i \leq n, \boldsymbol{x}_i \in A\}$. For the average (17) of the two estimators, the noiseless version is

$$\overline{f}_n^*(\boldsymbol{x}) = \frac{1}{2}\{\overline{f}_n^{(\text{max-min})}(\boldsymbol{x}) + \overline{f}_n^{(\text{min-max})}(\boldsymbol{x})\} \quad \forall \boldsymbol{x} \in V. \tag{58}$$

The functions in (57) and (58) can be viewed as estimation targets.

Our results can be summarized as follows. If we treat $\widehat{f}_n^{(\text{block})}(\boldsymbol{x}) - \overline{f}_n^*(\boldsymbol{x})$ as the estimation error and use $\overline{f}_n^*/\sigma$ to measure the range-to-noise ratio, all the theoretical results we have presented so far hold in the nonparametric regression model (56) for general f with the following adjustments of the error bounds $r_{q,\pm}(m)$ in (21) and (22):

$$
\begin{aligned}
&r_{q,+}(m) \ge \max\left\{\mathbb{E}\left[\max_{\boldsymbol{v}'\succeq\boldsymbol{v}}\left(\max_{\boldsymbol{u}\preceq\boldsymbol{x}}\sum_{\boldsymbol{x}_i\in[\boldsymbol{u},\boldsymbol{v}']}\frac{\varepsilon_i}{n_{\boldsymbol{u},\boldsymbol{v}'}}\right)_+^q\right] : n_{\boldsymbol{x},\boldsymbol{v}} = m, \boldsymbol{x}\preceq\boldsymbol{v} \text{ and } \boldsymbol{v}\in V_0\right\},\\
&r_{q,-}(m) \ge \max\left\{\mathbb{E}\left[\max_{\boldsymbol{u}'\preceq\boldsymbol{u}}\left(\min_{\boldsymbol{v}\succeq\boldsymbol{x}}\sum_{\boldsymbol{x}_i\in[\boldsymbol{u}',\boldsymbol{v}]}\frac{\varepsilon_i}{n_{\boldsymbol{u}',\boldsymbol{v}}}\right)_-^q\right] : n_{\boldsymbol{u},\boldsymbol{x}} = m, \boldsymbol{u}\preceq\boldsymbol{x} \text{ and } \boldsymbol{u}\in V_0\right\},
\end{aligned}
\tag{59}
$$

without changing the notation. Both $r_{q,\pm}(m)$ are still required to be nonincreasing functions of $m \in \mathbb{N}^+$. Accordingly, this leads to the following adjustment of the functions in (23):

$$
\begin{aligned}
m_{\boldsymbol{x},-} &= \max\{n_{\boldsymbol{u},\boldsymbol{x}} : \overline{f}^*_n(\boldsymbol{u}) \ge \overline{f}^*_n(\boldsymbol{x}) - r_{q,-}^{1/q}(n_{\boldsymbol{u},\boldsymbol{x}}), \boldsymbol{u}\preceq\boldsymbol{x} \text{ and } \boldsymbol{u}\in V_0\},\\
\boldsymbol{u}_{\boldsymbol{x}} &= \underset{\boldsymbol{u}\in V_0:\boldsymbol{u}\preceq\boldsymbol{x}}{\arg\max}\{n_{\boldsymbol{u},\boldsymbol{x}} : \overline{f}^*_n(\boldsymbol{u}) \ge \overline{f}^*_n(\boldsymbol{x}) - r_{q,-}^{1/q}(n_{\boldsymbol{u},\boldsymbol{x}})\},\\
m_{\boldsymbol{x}} &= m_{\boldsymbol{x},+} = \max\{n_{\boldsymbol{x},\boldsymbol{v}} : \overline{f}^*_n(\boldsymbol{v}) \le \overline{f}^*_n(\boldsymbol{x}) + r_{q,+}^{1/q}(n_{\boldsymbol{x},\boldsymbol{v}}), \boldsymbol{x}\preceq\boldsymbol{v} \text{ and } \boldsymbol{v}\in V_0\},\\
\boldsymbol{v}_{\boldsymbol{x}} &= \underset{\boldsymbol{v}\in V_0:\boldsymbol{x}\preceq\boldsymbol{v}}{\arg\max}\{n_{\boldsymbol{x},\boldsymbol{v}} : \overline{f}^*_n(\boldsymbol{v}) \le \overline{f}^*_n(\boldsymbol{x}) + r_{q,+}^{1/q}(n_{\boldsymbol{x},\boldsymbol{v}})\},
\end{aligned}
\tag{60}
$$

with the error bounds $r_{q,\pm}(m)$ in (59) and the estimation target $\overline{f}^*_n(\boldsymbol{x})$ in (57) or (58).

THEOREM 9. *Let $\widehat{f}_n^{(\text{block})}$ be as in (17), $\overline{f}^*_n$ as in (58), $r_{q,\pm}(m)$ as in (59) and $\ell_\pm(m)$ as in (24) with the $m_{\boldsymbol{x}_i,\pm}$ in (60). Then the error bounds (25) and (26) of Theorem 1 hold with f replaced by $\overline{f}^*_n$. Consequently, for the lattice design and under the $q \vee 2$ moment assumption on the noise $\{\varepsilon_i\}$, the error bounds in Theorems 2, 3, 4 and 5 hold with the same substitution. In particular, with f replaced by $\overline{f}^*_n$ and $\boldsymbol{f}_n$ by $\overline{\boldsymbol{f}}^*_n = (\overline{f}^*_n(\boldsymbol{x}_1), \dots, \overline{f}^*_n(\boldsymbol{x}_n))^T$, (36) holds with the same function $\widetilde{H}(t)$ when $\{\overline{f}^*_n(\boldsymbol{b}) - \overline{f}^*_n(\boldsymbol{a})\}/\sigma \le \Delta^*_n$, (41) and (43) hold when $\Delta(\overline{f}^*_n/n) \le \Delta^*_n$, (44) holds when $\overline{f}^*_n(\boldsymbol{a}_k) = \overline{f}^*_n(\boldsymbol{b}_k)$ with $V = \bigcup_{k=1}^K [\boldsymbol{a}_k, \boldsymbol{b}_k]$, and (45) and (46) hold when $\overline{f}^*_n(\boldsymbol{x})$ depends on only s of the d variables and $n_j \asymp n^{1/d}$ for all j. The above results also hold when $\{\widehat{f}_n^{(\text{block})}, \overline{f}^*_n\} = \{\widehat{f}_n^{(\text{max-min})}, \overline{f}_n^{(\text{max-min})}\}$ or $\{\widehat{f}_n^{(\text{block})}, \overline{f}^*_n\} = \{\widehat{f}_n^{(\text{min-max})}, \overline{f}_n^{(\text{min-max})}\}$.*

Theorem 9 asserts that $\widehat{\boldsymbol{f}}_n$ is close to $\overline{\boldsymbol{f}}^*_n$ in many ways when the isotonic condition on the unknown f is misspecified. However, the interpretation of this result is not as clear as the existing oracle inequality for the LSE as $\overline{\boldsymbol{f}}^*_n$ is not based on an optimality criterion.

4. Simulation results.

In this section, we report the results of several experiments in $d = 2$ and $d = 3$ to demonstrate the feasibility of the block estimators and to compare its estimation performance with the LSE. Among potentially many choices of the block estimator, we simply use the block max-min estimator as in (10). In six simulation settings, the block max-min estimator yields smaller average ℓ_2 losses than the LSE, with very small p-values in piecewise constant and variable selection settings. In a seventh setting, the LSE slightly outperforms the block max-min estimator but the difference is insignificant.

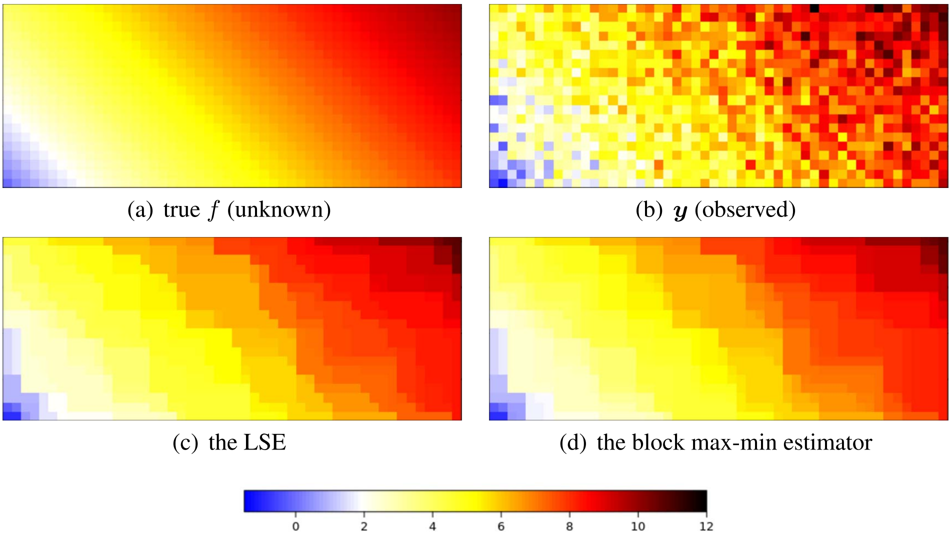
To compare the LSE and the block estimator, we carry out our experiments as follows. In each experiment, we generate one unknown $\boldsymbol{f}$, 5000 replications of $\boldsymbol{y}$ with standard Gaussian noise, find the LSE and the block max-min estimator for each $\boldsymbol{y}$ and compute the mean squared losses $\|\widehat{\boldsymbol{f}}_n - \boldsymbol{f}_n\|_2^2/n$ for both estimators. We therefore obtain 5000 simulated losses for each estimator and take the averages to approximate their mean squared risks.

We use quadratic programming to compute the LSE in our experiments. We would like to mention that fast algorithms for the LSE have been developed in the literature: Dykstra (1983), Kyng, Rao and Sachdeva (2015), Stout (2015), to name a few. We stick to quadratic programming as it provides somewhat more accurate results, although the difference seems small. The purpose of our experiment is to compare the risk of estimators, not the computational complexity of different algorithms. For the block max-min estimator, we use brute force which exhaustively calculates means over all blocks and finds the max-min value for each lattice point $\boldsymbol{x}$. We note again that the computation cost via brute force is of order n^3.

In $d = 2$, we consider isotonic regression with the $n_1 \times n_2$ lattice design $[\mathbf{1}, \boldsymbol{n}]$ with $n_1 = 50$ and $n_2 = 20$, so that the number of design points in total is $n = 1000$. In Experiment I, we consider the function $f(\boldsymbol{x}) = c(x_1 + x_2)^{2/3}$ (here and in the sequel, c is a constant such that $f(\boldsymbol{n}) = 10$ so that the range of f is about 10 on the lattice). As the region between two contours of this f cannot be efficiently represented by rectangular blocks, this example is not expected to favor the block estimator. In Experiment II, we split the lattice into 5×5 small blocks of size 10×4, randomly assign $1, \dots, 10$ to each small block, conditionally on the realizations satisfying the isotonic constraint. The adaptation of the LSE and the block max-min estimator to piecewise constant f is compared in this experiment. Lastly, we compare the adaptation of the two estimators to variable selection in Experiment III by setting $f(\boldsymbol{x}) = f_1(x_1) = c\log(x_1)$. See Figures 4, 5 and 6 for heat maps in Experiments I, II and III, respectively; each figure contains heat maps for the unknown f, one example of observed $\boldsymbol{y}$, the LSE and the block max-min estimator for this $\boldsymbol{y}$. Figure 7 provides boxplots of mean squared losses of both estimators in Experiments I, II and III.

In $d = 3$, we consider isotonic regression with $n_1 \times n_2 \times n_3$ lattice designs where $n_1 = n_2 = n_3 = 10$, so that the number of design points in total is also $n = 1000$. We choose the true mean functions in a similar manner to $d = 2$. In Experiment IV, we consider $f(\boldsymbol{x}) = c(x_1 + x_2 + x_3)^{2/3}$. In Experiment V, we randomly assign $1, \dots, 10$ to $2 \times 2 \times 5$ small blocks of size $5 \times 5 \times 2$ conditionally on the isotonic constraint. Lastly, the true mean function is $f(\boldsymbol{x}) = f_1(x_1) = c\log(x_1)$ in Experiment VI. See Figure 8 for boxplots of mean squared losses of both estimators in Experiments IV, V and VI.

Two basic statistics, mean and standard deviation of the losses of the LSE and the block max-min estimator and the loss difference of the two estimators are listed in Table 1, along with the two-sided p-value for the difference. In Experiment I and IV which are less favorable

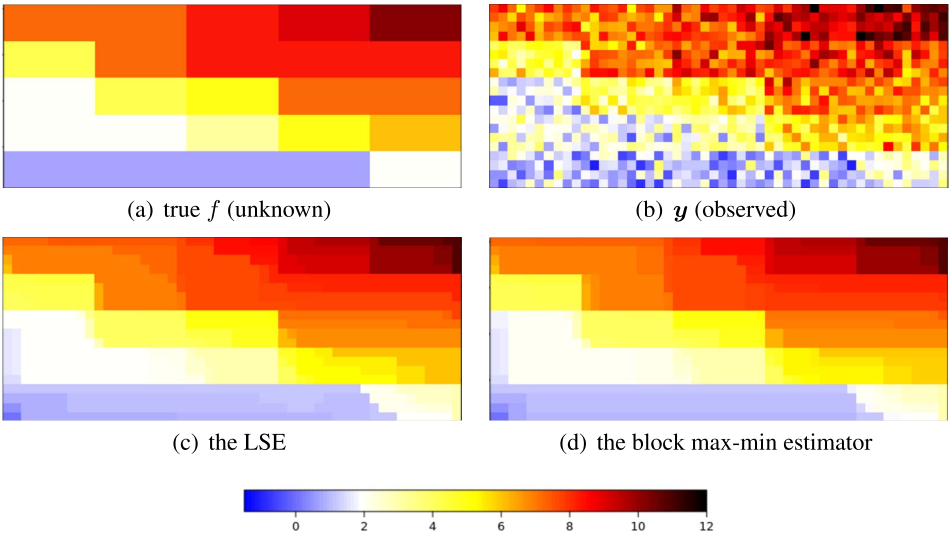

(a) true f (unknown)

(b) $\boldsymbol{y}$ (observed)

(c) the LSE

(d) the block max-min estimator

FIG. 4. *Heatmaps for the true $\boldsymbol{f}$, an observed $\boldsymbol{y}$ and its LSE and max-min estimate in Experiment* I.


(a) true f (unknown)

(b) $\boldsymbol{y}$ (observed)

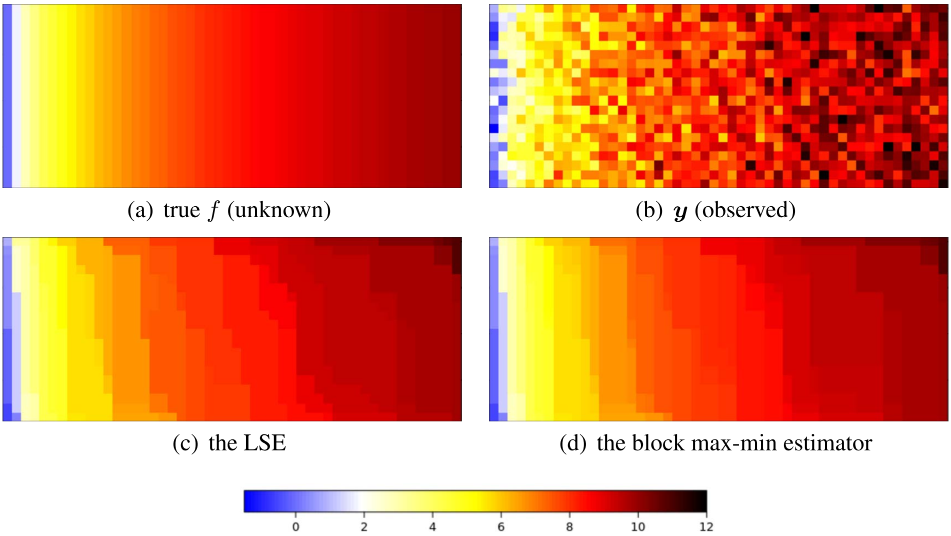
(c) the LSE

(d) the block max-min estimator

FIG. 5. *Heatmaps for the true piecewise-constant* $\boldsymbol{f}$, *an observed* $\mathbf{y}$, *and its LSE and max-min estimate in Experiment* II.

to the block estimator, the block estimator still yields slightly smaller risk, although the risk difference is insignificant (with p-values 0.6190 and 0.1600, resp.) In all other four experiments, the block max-min estimator significantly outperforms the LSE with p-values 0.0062 or smaller, supporting our theoretical analysis. It is worthwhile to mention that, although the risk values are incomparable due to different dimension d, we observe more significant difference in the mean squared losses between the LSE and the block max-min estimator in $d = 3$ than in $d = 2$, in view of the p-values and box plots. This observation coincide with Theorem 4 and its comparison to the existing risk bounds for the LSE.

We end this section with an example in which the LSE actually yields slightly smaller mean squared risk than the block max-min estimator. In Experiment VII, we consider the two-piece function $f(x_1, x_2) = I\{x_1/n_1 + x_2/n_2 \geq 1\}$ on an $n_1 \times n_2$ lattice. Same as in

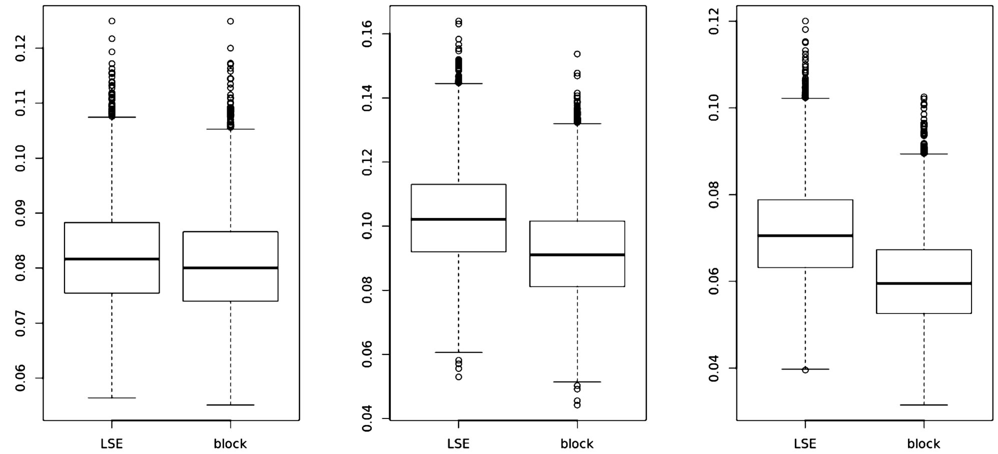

(a) true f (unknown)

(b) $\boldsymbol{y}$ (observed)

(c) the LSE

(d) the block max-min estimator

FIG. 6. *Heatmaps for the true* $\boldsymbol{f}$, *an observed* $\mathbf{y}$, *and its LSE and max-min estimate in Experiment* III.



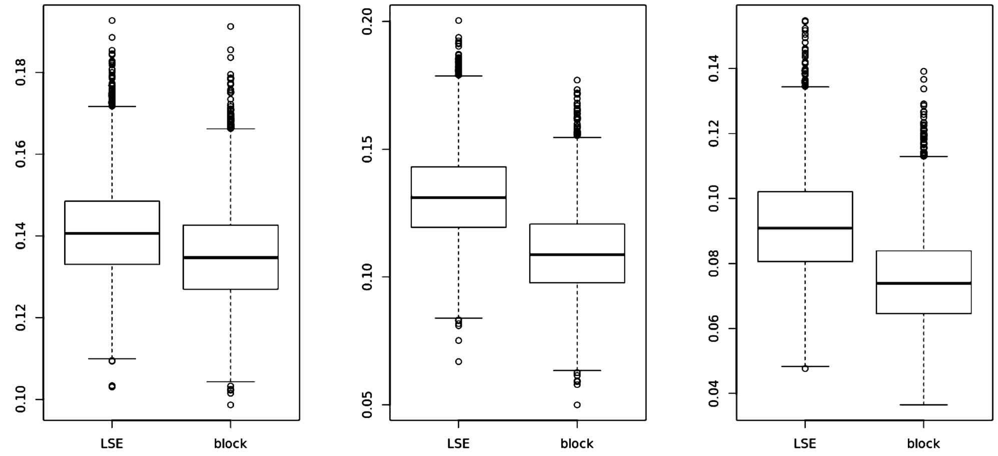
(a) Experiment I (b) Experiment II (c) Experiment III

FIG. 7. *Boxplots for the losses of LSE and block estimator in $d = 2$.*

Experiments I, II and III, we take $(n_1, n_2) = (50, 20)$ and add standard Gaussian noises to $f(x_1, x_2)$. See the heat maps in Figure 9.

We shall recall $\widehat{f}_n^{(\mathrm{lse})}(\boldsymbol{x}) = \overline{y}_{U \cap L}$ for some upper set U and lower set L. Suppose $x_1/n_1 + x_2/n_2 \geq 1$ so that $f(\boldsymbol{x}) = 1$, then the best level set $U \cap L$ for this design point is the upper red triangle in Figure 9(a). In contrast, as $\widehat{f}_n^{(\mathrm{block})}(\boldsymbol{x}) = \overline{y}_{[\boldsymbol{u},\boldsymbol{v}]}$ for some $\boldsymbol{u}$ and $\boldsymbol{v}$, the best possible block contains at most half design points of the upper triangle (when $\boldsymbol{u} = (n_1/2, n_2/2)$ and $\boldsymbol{v} = (n_1, n_2)$). Therefore, the variability of the block estimator at each design point may be larger than the LSE, resulting in a greater risk. Indeed, when we compare them on 5000 replications of $\boldsymbol{y}$ as in Experiments I–VI, the mean squared losses for the LSE has mean 0.0420 and standard deviation 0.0090, while for the block max-min estimator the mean is 0.0440 and the standard deviation is 0.0087. However, the difference is not significant as the mean and standard deviation for the loss difference are −0.0020 and 0.0040, and the two-sided p-value is 0.6163.

It would be difficult to characterize settings or general examples in which the LSE outperforms the block estimator. When we set $f(\boldsymbol{x}) = 0.5I\{x_1/n_1 + x_2/n_2 \geq 1\}$, the average



(a) Experiment IV (b) Experiment V (c) Experiment VI

FIG. 8. *Boxplots for the losses of LSE and block estimator in $d = 3$.*

TABLE 1

The mean and standard deviation (*s.d.*) *of the mean squared losses for the LSE and the block max-min estimator* (*block*), *and the mean, s.d. and two-sided p-value for the loss differences* (*diff* $=$ *loss of LSE* $-$ *loss of block estimator*)

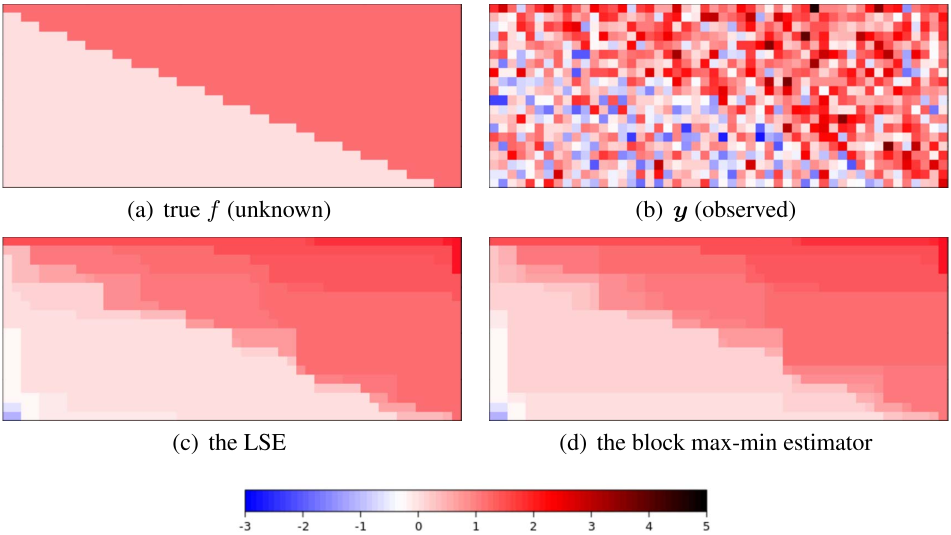
($d=2$)	Experiment I			Experiment II			Experiment III		
	LSE	block	diff	LSE	block	diff	LSE	block	diff
Mean	0.0822	0.0807	0.0016	0.1029	0.0918	0.0111	0.0713	0.0603	0.0110
s.d.	0.0096	0.0095	0.0031	0.0156	0.0149	0.0041	0.0115	0.0109	0.0033
p-value			0.6190			0.0062			0.0007
($d=3$)	Experiment IV			Experiment V			Experiment VI		
	LSE	block	diff	LSE	block	diff	LSE	block	diff
Mean	0.1412	0.1353	0.0059	0.1316	0.1096	0.0220	0.0917	0.0746	0.0170
s.d.	0.0119	0.0117	0.0042	0.0178	0.0169	0.0059	0.0160	0.0147	0.0045
p-value			0.1600			0.0002			0.0002

normalized ℓ_2 loss for the LSE is 0.0298, slightly greater than 0.0280 for the block max-min estimator, but the difference is insignificant as the two-sided p-value is 0.5568.

Acknowledgments. H. Deng's research is supported in part by NSF Grant DMS-1454817. C.-H. Zhang's research is supported in part by NSF Grants DMS-1513378, IIS-1407939, DMS-1721495, IIS-1741390 and CCF-1934924.

SUPPLEMENTARY MATERIAL

Supplement to "Isotonic regression in multidimensional spaces and graphs" (DOI: 10.1214/20-AOS1947SUPP; .pdf). This supplement contains proofs of all the theoretical results stated in the main body of the paper.



(a) true f (unknown)

(b) $\boldsymbol{y}$ (observed)

(c) the LSE

(d) the block max-min estimator

FIG. 9. *Heatmaps for the true two-piece function* $\boldsymbol{f}$, *an observed* $\mathbf{y}$ *and its LSE and max-min estimate.*

REFERENCES

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

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

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719 https://doi.org/10.1007/BF01199316

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

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

CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli* **24** 1072–1100. MR3706788 https://doi.org/10.3150/16-BEJ865

DENG, H. and ZHANG, C.-H. (2020). Supplement to "Isotonic regression in multi-dimensional spaces and graphs." https://doi.org/10.1214/20-AOS1947SUPP

DONOHO, D. L. (1990). Gelfand n-widths and the method of least squares. Preprint.

DUROT, C. (2007). On the $\mathbb{L}_p$-error of monotonicity constrained estimators. *Ann. Statist.* **35** 1080–1104. MR2341699 https://doi.org/10.1214/009053606000001497

DUROT, C. (2008). Monotone nonparametric regression with random design. *Math. Methods Statist.* **17** 327–341. MR2483461 https://doi.org/10.3103/S1066530708040042

DYKSTRA, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **78** 837–842. MR0727568

FOKIANOS, K., LEUCHT, A. and NEUMANN, M. H. (2017). On integrated L^1 convergence rate of an isotonic regression estimator for multivariate observations. Preprint. Available at arXiv:1710.04813.

GAO, C., HAN, F. and ZHANG, C.-H. (2017). Minimax risk bounds for piecewise constant models. Preprint. Available at arXiv:1705.06386.

GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153 (1957). MR0093415 https://doi.org/10.1080/03461238.1956.10414944

GROENEBOOM, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 539–555. Wadsworth, Belmont, CA. MR0822052

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

KYNG, R., RAO, A. and SACHDEVA, S. (2015). Fast, provable algorithms for isotonic regression in all l_p-norms. In *Advances in Neural Information Processing Systems* 2719–2727.

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

PRAKASA RAO, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36. MR0267677

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

STOUT, Q. F. (2015). Isotonic regression for multiple independent variables. *Algorithmica* **71** 450–470. MR3331888 https://doi.org/10.1007/s00453-013-9814-z

VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343 https://doi.org/10.1214/aos/1176347632

VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44. MR1212164 https://doi.org/10.1214/aos/1176349013

WANG, Y. (1996). The L_2 risk of an isotonic estimate. *Comm. Statist. Theory Methods* **25** 281–294. MR1379445 https://doi.org/10.1080/03610929608831695

WOODROOFE, M. and SUN, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when f is nonincreasing. *Statist. Sinica* **3** 501–515. MR1243398

YANG, F. and BARBER, R. F. (2019). Contraction and uniform convergence of isotonic regression. *Electron. J. Stat.* **13** 646–677. MR3914177 https://doi.org/10.1214/18-ejs1520

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