Annals of the Institute of Statistical Mathematics (2022) 74:489–514
https://doi.org/10.1007/s10463-021-00808-0

Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals

Alexander I. Jordan[1] · **Anja Mühlemann**[2] · **Johanna F. Ziegel**[2]

Received: 27 November 2020 / Revised: 25 June 2021 / Accepted: 13 July 2021 /
Published online: 3 September 2021

Abstract

In general, the solution to a regression problem is the minimizer of a given loss criterion and depends on the specified loss function. The nonparametric isotonic regression problem is special, in that optimal solutions can be found by solely specifying a functional. These solutions will then be minimizers under all loss functions simultaneously as long as the loss functions have the requested functional as the Bayes act. For the functional, the only requirement is that it can be defined via an identification function, with examples including the expectation, quantile, and expectile functionals. Generalizing classical results, we characterize the optimal solutions to the isotonic regression problem for identifiable functionals by rigorously treating these functionals as set-valued. The results hold in the case of totally or partially ordered explanatory variables. For total orders, we show that any solution resulting from the pool-adjacent-violators algorithm is optimal.

Keywords Order-restricted optimization problems · Partial order · Simultaneous optimality · Pool-adjacent-violators algorithm · Consistent loss functions

✉ Alexander I. Jordan
alexander.jordan@h-its.org

Anja Mühlemann
anja.muehlemann@stat.unibe.ch

Johanna F. Ziegel
johanna.ziegel@stat.unibe.ch

[1] Computational Statistics (CST) Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany

[2] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland

1 Introduction

Suppose that we have pairs of observations $(z_1, y_1), \ldots, (z_n, y_n)$ where we assume that y_i, $i = 1, \ldots, n$ are real-valued. The aim of isotonic regression is to fit an increasing function $\hat{g} : \{z_1, \ldots, z_n\} \to \mathbb{R}$ to these observations. The covariates $z_1, \ldots, z_n$ can take values in any set as long as it is equipped with a partial order which we denote by $\preceq$. Then, a function $g : \{z_1, \ldots, z_n\} \to \mathbb{R}$ is *increasing* if $z_i \preceq z_j$ implies that $g(z_i) \leq g(z_j)$.

As it is common in regression analysis, we aim to find an estimate $\hat{g}$ that minimizes the expected loss for some loss function $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$. If the function $\hat{g}$ is interpreted as an estimator of the conditional expectation of a random variable Y given Z, then a natural choice for L is the squared error loss $L(x, y) = (x - y)^2$. For $i \leq j$, let $\mathbb{E}_{i:j}$ denote the expectation with respect to the empirical distribution of $(z_i, y_i), \ldots, (z_j, y_j)$. Assuming that $z_1 < z_2 < \cdots < z_n$, the minimizer of the quadratic loss criterion

$$\mathbb{E}_{1:n}(g(Z) - Y)^2 \tag{1}$$

over all increasing functions g is given by

$$\hat{g}\big(z_\ell\big) = \min_{j \geq \ell} \max_{i \leq j} \mathbb{E}_{i:j} Y = \max_{i \leq \ell} \min_{j \geq i} \mathbb{E}_{i:j} Y, \quad \ell = 1, \ldots, n, \tag{2}$$

see Barlow et al. (1972, eq. (1.9)–(1.13)). The solution $\hat{g}$ can be computed efficiently using the so-called pool-adjacent-violators (PAV) algorithm. These results were developed in the 1950s by several parties independently; see Ayer et al. (1955); Bartholomew (1959a, 1959b); Brunk (1955); van Eeden (1958); Miles (1959).

It turns out that the solution given at (2) is also the unique minimizer of the Bregman loss criterion

$$\mathbb{E}_{1:n} L(g(Z), Y), \tag{3}$$

where the squared error loss in (1) has been replaced by a Bregman loss function $L = L_\phi$ (Barlow et al. 1972, Theorem 1.10). That is,

$$L_\phi(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x),$$

where ϕ is a convex function with subgradient ϕ'. Savage (1971) found that the Bregman class comprises all loss functions L where the expectation functional minimizes the expected loss, i.e.,

$$\mathbb{E}_P Y = \arg\min_x \mathbb{E}_P L(x, Y),$$

where Y is a random variable with distribution P. Due to this property, any loss function in the Bregman class is also referred to as a consistent loss function for the expectation functional (Gneiting 2011).

In summary, the increasing regression function at (2) is simultaneously optimal with respect to all consistent loss functions for the expectation. This robustness with respect to the choice of loss function means that the solution to the regression

problem is determined by the choice of the expectation as the target functional. We will see that the same holds for other functionals. As such, in nonparametric isotonic regression we can replace the task of choosing a loss function with the task of choosing a suitable target functional.

This remarkable result is particularly beneficial in scenarios where a single relevant loss function cannot easily be identified. For example, institutions such as central banks or weather services provide analyses and forecasts that drive individual decision making in a heterogeneous group of users. In these circumstances, determining a unifying loss function is hardly trivial. However, publishing results for the expectation and for various quantile levels is certainly feasible.

The simultaneous-optimality result for nonparametric isotonic regression is in stark contrast to the optimality behavior of parametric models for increasing regression functions. Suppose that $\{g_\theta : \theta \in \Theta\}, \Theta \subseteq \mathbb{R}^d$ is a parametric model of increasing functions g_θ. Then, the optimal parameters with respect to the Bregman-loss criterion (3) generally vary (substantially) depending on the chosen loss function (Patton 2020). Consistency of the loss function merely ensures that the true parameter value of a correctly specified model minimizes the Bregman-loss criterion on the population level. Interestingly, simultaneous optimality with respect to all consistent loss functions generally also breaks down if one weakens the isotonicity constraint of the regression function to a unimodality constraint; see Sect. 3.

In this paper, we generalize the result of Barlow et al. (1972, Theorem 1.10) in several directions. First, instead of the expectation functional, we consider general functionals T that are given by an identification function $V(x, y)$ as defined in Definition 1. Second, in the case of set-valued functionals, we give a complete characterization of all possible solutions. Third, we demonstrate that modified min-max and max-min solutions as in (2) continue to hold for general partial orders on the covariates and act as lower and upper bounds to any solution.

An identification function is an increasing function that weighs negative values in the case of underestimation against positive values in the case of overestimation, with an optimal expected value of zero. The corresponding functional T then maps to the optimizing value of the argument (or to the set of optimizing values). If there is always a unique optimizing value, we say that the functional is of singleton type, and otherwise it is of interval type. Prime examples of functionals that are of singleton type include the expectation functional, expectiles (Newey and Powell 1987), or ratios of expectations. The solution for these functionals is unique, so that our results offer no new insight to those by Robertson and Wright (1980) apart from a different method of proof. Functionals that are of interval type include the important case of quantiles, including the median, which have also been treated in Robertson and Wright (1973, 1980), but not in the interpretation as set-valued functionals. Predefining a global scheme for reducing the median interval to a single point (e.g., some convex combination of lower and upper functional value) inevitably restricts the possible solutions to the isotonic regression problem.

In contrast to previous work, we treat all functionals as set-valued. In Sect. 4, we give explicit solutions for the lower and upper bound of the isotonic regression problem in the context of partial orders. The method of proof for these results is fundamentally different from the approach of Barlow et al. (1972, Theorem 1.10) or

Robertson and Wright (1980), and in contrast to the latter comes with an immediate construction principle for loss functions. Our method relies on the mixture or Choquet representations of consistent loss functions, introduced by Ehm et al. (2016) for the quantile and expectile functionals. Given the identification function $V(x, y)$ for the functional T, a one-parameter family of elementary loss functions that are consistent for the functional T can be readily defined,

$$S_\eta(x, y) = (\mathbb{1}\{\eta \le x\} - \mathbb{1}\{\eta \le y\})V(\eta, y),$$

where $\eta \in \mathbb{R}$. If T is a quantile, an expectile, or a ratio of expectations, then

$$\mathscr{S} = \left\{ \int_{\mathbb{R}} S_\eta(x, y)\, \mathrm{d}H(\eta) \,:\, H \text{ is a nonnegative measure on } \mathbb{R} \right\}, \tag{4}$$

comprises all consistent loss functions for T subject to standard conditions, and if $V(x, y) = x - y$ is the identification function of the expectation, then the class $\mathscr{S}$ is the class of Bregman loss functions; see Ehm et al. (2016); Gneiting (2011). In fact, optimality of an isotonic solution with respect to the criterion (3) for $L = S_\eta$ for some $\eta \in \mathbb{R}$ corresponds to finding a solution with optimal superlevel set $\{g \ge \eta\}$. Considering an isotonicity constraint as a constraint on admissible superlevel sets of the regression function relates to the work of Polonik (1998) in the context of density estimation.

For any functional T and corresponding consistent loss function L from the class $\mathscr{S}$, Theorem 1 states that the optimal isotonic solution to the criterion (3) is bounded below by a min-max formula and bounded above by a max-min formula as in (2) with the expectation replaced by the lower and upper functional values under T, respectively. In Proposition 5, we show that the min-max or max-min solution is simultaneously optimal with respect to all elementary loss functions for T, and hence with respect to the entire class $\mathscr{S}$. Propositions 6–8 characterize the optimal solutions by refinement of other optimal solutions. Our method of proof also leads to a transparent proof of the validity of the PAV algorithm in Sect. 4.2.

The left panel of Fig. 1 illustrates the pointwise bounds given in Theorem 1 for the median functional in a constructed data example with totally ordered covariates. The right panel illustrates the full range of optimal solutions as given by Propositions 5–8. These propositions identify all optimal superlevel sets and thereby also the regions where an optimal solution is necessarily constant (shown in darkgrey), interspersed with regions where the only constraint is that isotonicity has to be satisfied (lightgrey). As examples we show one optimal solution that linearly interpolates the midpoints of the bounds from Theorem 1 (red), and another solution that minimizes the average slope subject to continuity (blue). Note that the latter solution has 5 constant pieces, which is impossible for a fixed convex combination of the bounds from Theorem 1.

The results in Robertson and Wright (1980) hold for a large class of functionals and for partial orders on the covariates. However, the generality of their results is limited by treating potentially set-valued functionals as maps to single values. The solutions that arise from Proposition 5 in combination with Corollary 3 should be recoverable in the framework of Robertson and Wright (1980), which is in general

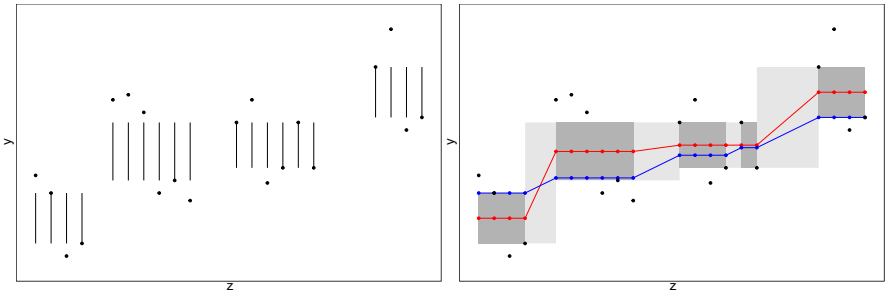

Fig. 1 Solutions in isotonic median regression. The left panel shows the pointwise bounds from Theorem 1. The right panel illustrates the full range of optimal solutions, which are necessarily constant in some regions (darkgrey) and can be chosen freely inbetween (lightgrey), subject to satisfying isotonicity. Two examples of optimal solutions are shown. One is based on Corollary 3 and interpolates linearly between the midpoints of the pointwise bounds (red), and one is based on Proposition 8 and minimizes the average slope (blue)

not the case for the solutions by Propositions 7–8. Also, the minimal and maximal solutions of Proposition 5 have not formally been identified as actual bounds, as they are in Theorem 1. To the best of our knowledge, the literature following Robertson and Wright (1980) is void of further results that characterize the solutions to the isotonic regression problem, or any investigations into the effect of the choice of loss function among options sharing the same Bayes act.

Recently and independently of our work, Mösching and Dümbgen (2020) derived a similar result of min-max and max-min formulas as lower and upper bounds for optimal isotonic solutions in the context of set-valued minimizers of convex and coercive loss functions. In contrast to their work, we do not require loss functions to be convex and coercive. Instead we focus on their consistency for a specific functional. Brümmer and Du Preez (2013) rediscover the result of Barlow et al. (1972) that the PAV algorithm leads to a simultaneously optimal solution for all proper scoring rules in the context of binary events – a special class of loss functions that are consistent for the expectation functional.

A comprehensive overview on isotonic regression is given in the monograph Groeneboom and Jongbloed (2014). Also, Guntuboyina and Sen (2018) review risk bounds, asymptotic theory, and algorithms in common nonparametric shape-restricted regression problems in the context of least squares optimization. Among the most recent developments on algorithms for isotonic regression with partially ordered covariates, Kyng et al. (2015) and Stout (2015) provide fast algorithms for isotone regression under different loss functions using the representation of a partial order as a directed acyclic graph. Recent advances on asymptotic theory for isotonic regression include Han et al. (2019), giving rates for least squares isotonic regression on the unit cube of arbitrary dimension, and Bellec (2018), considering isotonic, unimodal, and convex regression in the context of total orders. Another recent interest is the regularization of isotonic regression on multiple variables with Luss and Rosset (2017) proposing a method via range restriction on the solution to the regression problem.

The paper provides a first fully rigorous treatment of isotonic regression for set-valued identifiable functionals including the important special case of quantiles. We treat total orders as well as partial orders. In both cases, a complete characterization of all solutions to the isotonic regression problem has been lacking in the literature. Theorem 1 identifies bounds on the solutions. Propositions 7–8 only bear relevance when the functional is of interval type, but they identify the conditions that lead to those optimal solutions which cannot occur when functionals only map to single real values.

2 Functionals and consistent loss functions

We start with the definition of a functional via an identification function.

Definition 1 A function $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is called an *identification function* if $V(\cdot, y)$ is increasing and left-continuous for all $y \in \mathbb{R}$. Then, for any finite and nonnegative measure P on $\mathbb{R}$ such that $V(x, \cdot)$ is quasi-integrable for all $x \in \mathbb{R}$, we define the *functional* T induced by an identification function V as

$$T(P) = \left[T_P^-, T_P^+\right] \subseteq [-\infty, +\infty] = \bar{\mathbb{R}},$$

where the lower and upper bounds are given by

$$T_P^- = \sup\{x : V(x, P) < 0\} \quad \text{and} \quad T_P^+ = \inf\{x : V(x, P) > 0\},$$

using the notation $V(x, P) = \int_{-\infty}^{\infty} V(x, y)\, \mathrm{d}P(y)$.

Defining functionals for any finite and nonnegative measure, as opposed to merely probability distributions, is a minor detail that simplifies notation when joining and intersecting data subsets. Except in the case of the null measure, any finite and nonnegative measure can be replaced with a corresponding probability distribution, without any change to the functional values.

Note that T_P^- can take the value $-\infty$, and T_P^+ can take the value $+\infty$. In the subsequent results, we repeatedly refer to the smallest or largest element of a finite set where one of the elements could be $\pm\infty$. We still write min and max of the set but this quantity could be $\pm\infty$.

Definition 2 A functional T is called a *functional of singleton type* if $T(P)$ is a singleton whenever P is not the null measure. Otherwise, T is called a *functional of interval type*.

Table 1 summarizes common functionals and their respective identification functions, and Example 1 explains two options in more detail.

Example 1 Let $\alpha, \tau \in (0, 1)$, and let P denote a probability distribution.

Table 1 Selection of functionals and their respective identification functions. The parameters satisfy $\alpha, \tau \in (0,1)$, $p > 1$ and $\delta > 0$, and $u : I \to \mathbb{R}$ and $w : I \to (0,\infty)$ are measurable functions on an interval $I \subseteq \mathbb{R}$. The functionals "ℓ_p minimizer" and "Huber minimizer" map to the intervals of values minimizing the ℓ_p loss and the Huber loss (Huber 1964), respectively

Functional	Identification function	Type
Median	$V(x,y) = \mathbb{1}\{x > y\} - 1/2$	interval
Mean	$V(x,y) = x - y$	singleton
2nd Moment	$V(x,y) = x - y^2$	singleton
α-Quantile	$V(x,y) = \mathbb{1}\{x > y\} - \alpha$	interval
τ-Expectile	$V(x,y) = 2\lvert\mathbb{1}\{x > y\} - \tau\rvert(x - y)$	singleton
Ratio $\mathbb{E}_P(u(Y))/\mathbb{E}_P(w(Y))$	$V(x,y) = xw(y) - u(y)$	singleton
ℓ_p minimizer	$V(x,y) = \operatorname{sign}(x-y)\lvert x-y\rvert^{p-1}$	singleton
Huber minimizer	$V(x,y) = \operatorname{sign}(x-y)\min(\lvert x-y\rvert, \delta)$	interval

(a) Consider the identification function $V(x,y) = \mathbb{1}\{x > y\} - \alpha$, then $V(x,P) = P(Y < x) - \alpha$, and the interval of all α-quantiles of P,

$$T(P) = \left[\sup\{x : P(Y < x) < \alpha\}, \inf\{x : P(Y < x) > \alpha\}\right],$$

is potentially of positive length.

(b) If P has a finite first moment, the identification function $V(x,y) = 2\lvert\mathbb{1}\{x > y\} - \tau\rvert(x - y)$ leads to

$$V(x,P) = 2(1-\tau)\int_{-\infty}^{x}(x-y)\,\mathrm{d}P(y) + 2\tau\int_{x}^{\infty}(x-y)\,\mathrm{d}P(y),$$

which is strictly increasing and continuous in its first argument. Hence, there exists a unique solution in x for the equation $V(x,P) = 0$, and we call that solution the τ-expectile $e_\tau(P)$. In particular, for $\tau = \frac{1}{2}$ we obtain $V(x,y) = x - y$ and thus $T(P) = \{\mathbb{E}_P(Y)\}$.

In the later proofs, we use three implications of Definition 1 repeatedly to establish order relationships between the variable in the first argument of V and the functional of an empirical distribution. To facilitate reference, we note these statements explicitly.

Corollary 1 *Let V be an identification function inducing the functional T, and P be a finite and nonnegative measure on $\mathbb{R}$. Then,*

$$\begin{aligned} V(\eta,P) = 0 &\implies \eta \in T(P), \\ V(\eta,P) > 0 &\implies \eta > \sup T(P) = T_P^+, \\ V(\eta,P) < 0 &\implies \eta \le \inf T(P) = T_P^-. \end{aligned}$$

Lemma 1 shows that a version of the Cauchy mean value property holds for any functional that we consider in this paper. The same can be shown for the original

version used to define functionals in Robertson and Wright (1980). It is unclear whether a functional that satisfies the Cauchy mean value property needs to be identifiable.

Lemma 1 *Let P, Q be finite and nonnegative measures on $\mathbb{R}$. Then,*

$$\min\left\{T_P^-, T_Q^+\right\} \leq T_{P+Q}^- \leq T_{P+Q}^+ \leq \max\left\{T_P^-, T_Q^+\right\}.$$

Proof The statement follows from Definition 1. The second inequality is trivial. For the first inequality, and $x < \min\{T_P^-, T_Q^+\}$, we have $V(x, P) < 0$ and $V(x, Q) \leq 0$, hence $V(x, P + Q) < 0$. A similar argument applies to the third inequality. □

The definition of a functional in terms of an identification function comes with a straightforward construction principle for large classes of loss functions. In a nutshell, a continuous oriented identification function defines a functional via its unique root in the first argument, a first-order condition. By integration, corresponding loss functions inherit the consistency for the functional, i.e., the minimum expected loss is attained by any member in $T(P)$. The loss functions defined in Proposition 1 are the most basic, in the sense that they are a result of integration with respect to the Dirac measure at a given threshold $\eta \in \mathbb{R}$. A similar result has also been discussed in Dawid (2016) and Ziegel (2016).

Proposition 1 *Let V be an identification function, T be the induced functional, and $\eta \in \mathbb{R}$. Then the elementary loss function $S_\eta : \bar{\mathbb{R}} \times \mathbb{R} \to \mathbb{R}$ given by*

$$S_\eta(x, y) = (\mathbb{1}\{\eta \leq x\} - \mathbb{1}\{\eta \leq y\})V(\eta, y)$$

is consistent for T relative to the class $\mathscr{P}$ of probability distributions such that $V(\eta, \cdot)$ is quasi-integrable. That is,

$$\mathbb{E}_P S_\eta(t, Y) \leq \mathbb{E}_P S_\eta(x, Y)$$

for all $P \in \mathscr{P}$, all $t \in T(P)$ and all $x \in \bar{\mathbb{R}}$.

Proof Let

$$d(\eta) = \mathbb{E}_P S_\eta(t, Y) - \mathbb{E}_P S_\eta(x, Y) = (\mathbb{1}\{\eta \leq t\} - \mathbb{1}\{\eta \leq x\})V(\eta, P).$$

If $V(\eta, P) = 0$ then $d(\eta) = 0$. If $V(\eta, P) < 0$ it follows from Corollary 1 that $\eta \leq t$ and therefore $d(\eta) \leq 0$. Similary, if $V(\eta, P) > 0$ it follows that $\eta > t$ and therefore $d(\eta) \leq 0$. □

As an immediate consequence of the consistency of elementary loss functions for the functional T, we have that all loss functions in the class $\mathscr{S}$ defined at (4) are also consistent for the functional T. This result exemplifies an important line of reasoning used multiple times in this paper: A property of S_η that holds for all $\eta \in \mathbb{R}$ translates to the class $\mathscr{S}$.

Examples of members of the class $\mathscr{S}$ for the expectation functional, i.e., $V(x, y) = x - y$, are given in Table 2. While these examples are differentiable convex losses and therefore already covered in the literature (Luss and Rosset 2014), the analysis in this paper also holds for the absolute loss, a nondifferentiable convex loss that is recovered when choosing $V(x, y) = \mathbb{1}\{x > y\} - 1/2$ and $\mathrm{d}H(\eta) = \mathrm{d}2\eta$. And even the elementary loss functions themselves bear relevance to fundamental decision problems in practice (Ehm et al. 2016). For the expectation functional, the elementary losses are nondifferentiable and convex, but describe the scenario of investing a fixed sum η for an unknown future profit or loss. For quantiles, the losses are not even convex, but describe the scenario of a bet on whether or not the outcome y will exceed the threshold η, with a fixed payoff ratio. Similar betting interpretations of elementary loss functions in the context of isotonic regression are an interesting open question.

While loss functions with properties such as convexity or differentiability are often necessary in optimization problems for estimation, consumers of predictions regularly face decision problems with simpler loss structures. The results in this paper show that a distinction of preferences for technical implementation and forecast consumption is unnecessary in nonparametric isotonic regression.

3 Simultaneous optimality

Consider a distribution P for a random vector $(Z, Y) \in \mathscr{Z} \times \mathbb{R}$. We aim to minimize the criterion

$$\mathbb{E}_P S_\eta(g(Z), Y) \quad \text{for all } \eta \in \mathbb{R}, \tag{5}$$

over a family of regression functions $g : \mathscr{Z} \to \mathbb{R}$, and call a solution $\hat{g}$ simultaneously optimal since it minimizes the expected score with respect to all scoring functions in the class $\mathscr{S}$ at (4), simultaneously. Condition (5) is equivalent to minimizing $\mathbb{E}_P \mathbb{1}\{\eta \le g(Z)\} V(\eta, Y)$ for all $\eta \in \mathbb{R}$. The results in this paper rely on this reformulation and the implication that regression functions are characterized by superlevel sets of the form $\{z \in \mathscr{Z} : \hat{g}(z) \ge \eta\}$, $\eta \in \mathbb{R}$. The structure of the set of admissible superlevel sets is crucial for the existence of a simultaneously optimal regression function.

Table 2 Commonly used loss functions that are consistent for the mean functional. For an interval $I \subseteq \mathbb{R}$, a Bregman loss is induced by a convex function $\phi : I \to \mathbb{R}$ with subgradient ϕ'. See Patton (2011, 2020) for the QLIKE loss and the exponential Bregman loss, respectively

Name	Mixing measure $H((\eta_1, \eta_2]) =$	Loss function $L(x, y) =$	Domain
Bregman loss	$\phi'(\eta_2) - \phi'(\eta_1)$	$\phi(y) - \phi(x) - \phi'(x)(y - x)$	I
Squared error	$\eta_2 - \eta_1$	$(x - y)^2$	$\mathbb{R}$
Exponential Bregman	$\exp(\eta_2) - \exp(\eta_1)$	$\exp(y) - \exp(x) - \exp(x)(y - x)$	$\mathbb{R}$
QLIKE loss	$-1/\eta_2 + 1/\eta_1$	$y/x - \log(y/x) - 1$	$(0, \infty)$

In fact, it is a rare property in regression methods, that the solution does not depend on the loss function when considering a large class such as $\mathscr{S}$. As recently demonstrated, the optimal parameters with respect to the Bregman-loss criterion (3) of a parametric model $\{g_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$ of increasing functions g_θ generally vary depending on the chosen loss function (Patton 2020). Before proving the simultaneous-optimality result for nonparametric isotonic regression in Sect. 4, we highlight the fragility of simultaneous optimality by demonstrating that it fails to hold for only slightly adapted shape constraints.

Unimodality is a shape constraint closely related to isotonicity. Given a predetermined mode, unimodality is even equivalent to isotonicity, when order relationships are defined suitably. For example, a total order on a finite set becomes a partial order consisting of two separate total orders merging in the predetermined mode, when reframing unimodality as isotonicity. Then, the problem becomes one of reconciling two isotonicity constraints. However, we will now see that simultaneous optimality under the unimodality constraint is in general unattainable when the location of the mode is not predetermined.

Example 2 Suppose that we have observations $(z_1, y_1), \dots, (z_4, y_4)$ with $z_1 < \cdots < z_4$ and $(y_1, \dots, y_4) = (9, 9, 0, 10)$, and let P denote the corresponding empirical distribution. We choose the expectation functional as the regression target, and for each potential mode $m_i = z_i$, $i = 1, \dots, 4$, we aim to find a function $\hat{g}_i : \{z_1, \dots, z_4\} \to \mathbb{R}$ that is optimal for any consistent loss function for the expectation functional. To this end, we reframe unimodality given a predetermined mode as isotonicity. The existence and the uniqueness of an optimal isotonic solution for a functional of singleton type is shown in Sect. 4.

Using the PAV algorithm, the functions $\hat{g}_1$ and $\hat{g}_4$ are easy to find, as the order on the z_i is reversed or remains unchanged, respectively, when reframing the unimodality constraint as isotonicity. We refer to Sect. 4.2 and extant literature for a description of the algorithm. To find $\hat{g}_3$, we consider the partial order given by the totally ordered subsets $z_1 < z_2 < z_3$ and $z_3 > z_4$, and argue with superlevel sets of the form $\{z : \hat{g}_3(z) \geq \eta\}$, $\eta \in \mathbb{R}$. Since z_4 corresponds to the largest response in the data set, y_4, and z_3 needs to be in every nonempty superlevel set, we have $\hat{g}_3(z_3) = \hat{g}_3(z_4)$. Therefore, z_4 also lies in any nonempty superlevel set of $\hat{g}_3$, and in satisfying the isotonic relationship on $z_1 < z_2 < z_3$, we find that the only nonempty superlevel set must be $\{z_1, \dots, z_4\}$, corresponding to levels $\eta \leq \frac{1}{4}\sum_{i=1}^4 y_i = 7$. Similarly, in order to find $\hat{g}_2$ as the isotonic solution subject to $z_1 < z_2$ and $z_2 > z_3 > z_4$, we again have $\hat{g}_2(z_3) = \hat{g}_2(z_4)$ since y_4 is the largest response. As $\frac{1}{2}\sum_{i=3}^4 y_i < y_2 = y_1$, isotonicity is established, and the only nonempty superlevel sets are $\{z_1, z_2\}$ and $\{z_1, \dots, z_4\}$, corresponding to levels $\eta \in (5, 9]$ and $\eta \leq 5$, respectively. Coincidentally, $\hat{g}_2 = \hat{g}_1$.

The left panel of Fig. 2 shows the regression functions, and the right panel shows the expected score at (5) as a function of $\eta \in \mathbb{R}$. None of the three potential solutions minimizes the expected score for all η, and therefore, a simultaneously optimal solution does not exist in this example. This visual method of comparing forecasts is called a Murphy diagram (Ehm et al. 2016).

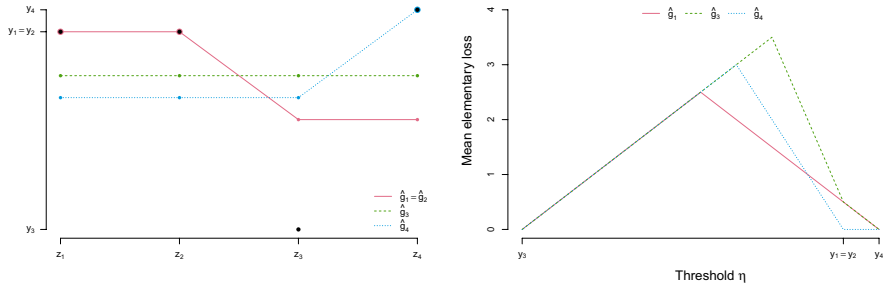

Fig. 2 Unimodal Regression and Murphy Diagram. For a data example with observations $(z_1, 9), (z_2, 9), (z_3, 0), (z_4, 10)$, the left panel shows the regression functions $\hat{g}_1, \ldots, \hat{g}_4$ corresponding to modes $z_1, \ldots, z_4$. The black dots display the observations. The right panel shows the mean elementary losses of the regression functions against the parameter $\eta \in \mathbb{R}$. No single function exhibits the smallest mean elementary loss for all values of η, simultaneously

In unimodal regression, a simultaneously optimal solution may but need not exist. This agrees with our findings in Sect. 4 because the set of admissible superlevel sets under a unimodality shape constraint is not closed under union and intersection. Indeed, in Example 2 the sets $\{z_1\}$ and $\{z_4\}$ are admissible superlevel sets, while the union $\{z_1, z_4\}$ is not admissible because it implies bimodality.

4 Results on isotonic regression

We solve the isotonic regression problem considering a distribution P for a random vector $(Z, Y) \in \mathscr{Z} \times \mathbb{R}$, where $\mathscr{Z}$ is a finite partially ordered set. The distribution P may, but need not, be an empirical distribution with finite support. Analogously to (5), we aim to minimize the criterion

$$\mathbb{E}_P S_\eta(g(Z), Y) \quad \text{for all } \eta \in \mathbb{R}, \tag{6}$$

over all increasing functions $g : \mathscr{Z} \to \bar{\mathbb{R}}$. We call any minimizer of (6) a solution to the isotonic regression problem.

Reformulation of condition (6) as minimizing $\mathbb{E}_P \mathbb{1}\{\eta \le g(Z)\} V(\eta, Y)$ for all $\eta \in \mathbb{R}$ reveals that we can specify a solution to the isotonic regression problem by finding a path through minimizing upper sets $\{z \in \mathscr{Z} : \hat{g}(z) \ge \eta\}$. These upper sets are denoted by $x \in \mathscr{X} \subseteq \mathscr{P}(\mathscr{Z})$, where $\mathscr{P}$ denotes the power set. The set $\mathscr{X}$ consists of all admissible superlevel sets for an increasing function g imposed by the partial order on $\mathscr{Z}$. A set $x \in \mathscr{X}$ is characterized by the property that if $z \in x$ and $z \preceq z'$, then $z' \in x$. This implies that $\mathscr{X}$ is a finite lattice, that is, it is closed under union and intersection and contains $\mathscr{Z}$ and the empty set. We will see, that as η increases, ξ follows one of the totally ordered paths through the lattice. In Fig. 3, the direction of movement as η increases is illustrated by arrows. In the special case of a total order, $z_1 \prec \cdots \prec z_n$, there is only one possible path along upper sets of the form $\{z_i, \ldots, z_n\}$, $i = 1, \ldots, n$, ending up at the empty set.

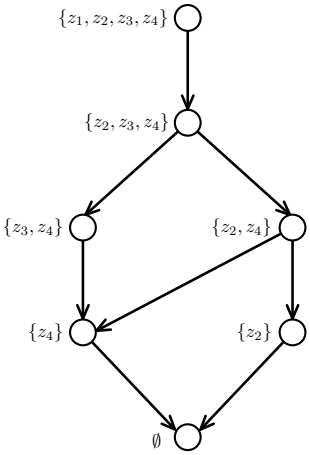

Fig. 3 Moving through the lattice $\mathscr{X}$. The display shows possible paths through $\mathscr{X}$ based on the partial order on $\mathscr{Z} = \{z_1, z_2, z_3, z_4\}$ given by $z_1 \prec z_2$ and $z_1 \prec z_3 \prec z_4$. The arrows indicate the direction of moving through the lattice $\mathscr{X}$ as η increases

The path is given by a function $\xi : \mathbb{R} \to \mathscr{X}$, that maps η to an upper set x of $\mathscr{Z}$ that minimizes

$$s_x(\eta) = v_x(\eta) = V(\eta, P_x) = \int_{x \times \mathbb{R}} V(\eta, y)\, P(\mathrm{d}z, \mathrm{d}y), \tag{7}$$

where $P_x(A) = P((x \times \mathbb{R}) \cap A)$ for any $A \in \mathscr{P}(\mathscr{Z}) \otimes \mathscr{B}(\mathbb{R})$, where $\mathscr{B}(\mathbb{R})$ denotes the Borel σ-algebra on $\mathbb{R}$. In this notation, s_x is only defined for $x \in \mathscr{X}$, whereas v_x and P_x are defined for any $x \in \mathscr{P}(\mathscr{Z})$. As in Definition 1 and Proposition 1, we assume that $V(\eta, P_x)$ exists for all $\eta \in \mathbb{R}$ and $x \in \mathscr{P}(\mathscr{Z})$. For the bounds of the conditional functional, we write $T_x^- = T_{P_x}^- = \inf T(P_x)$ and $T_x^+ = T_{P_x}^+ = \sup T(P_x)$. Finally, let $X(\eta)$ denote the set of superlevel sets $x \in \mathscr{X}$ minimizing $s_x(\eta)$ at (7). Since $\mathscr{P}(\mathscr{Z})$ is finite, such a minimizer always exists.

For a total order, upper sets $\{z_i, \ldots, z_n\}$ can be parameterized by the index of the smallest element, with the index $n+1$ for the empty set. Then we can redefine the object of minimization in (7) as

$$s_i(\eta) = \sum_{\ell=i}^{n} V(\eta, y_\ell).$$

This index search needs to be conducted for every $\eta \in \mathbb{R}$ separately. In Fig. 4, we give an example for 6 data points. The example illustrates how the values $\hat{g}(z_\ell)$, $\ell = 1, \ldots, 6$, can be determined from the epigraph of the function $\eta \mapsto \min \xi(\eta)$. The function ι maps η to the smallest index of the elements in $\xi(\eta)$. In a nutshell, for a total order, we find the generalized inverse to an optimal solution.

The following proposition formalizes that statement in the general context, assuming the existence of a decreasing function $\xi : \mathbb{R} \to \mathscr{X}$ in the sense that for $\eta' > \eta$ it holds that $\xi(\eta') \subseteq \xi(\eta)$, while satisfying $\xi(\eta) \in X(\eta)$ for all $\eta \in \mathbb{R}$. Before showing the existence of such a function ξ in Lemma 3, we elucidate the one-to-one correspondence to the solutions $\hat{g}$ of the isotonic regression problem at (6).

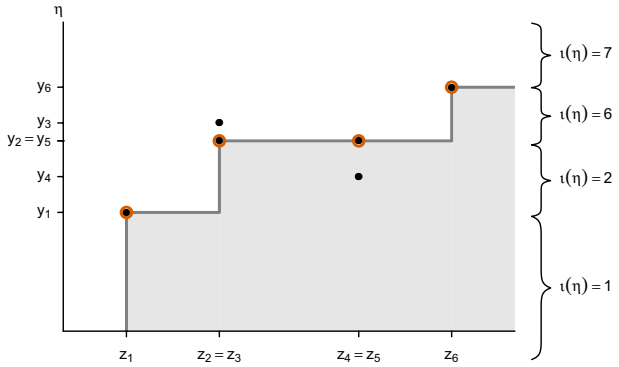

Fig. 4 Graph of $\hat{g}$. For a sample of 6 data points with a totally ordered covariate set $\mathscr{Z}$, the values of $\hat{g}(z)$ for $z = z_1, \ldots, z_6$ are shown in red. The epigraph of the function $\eta \mapsto \min \xi(\eta) = z_{\iota(\eta)}$ is shown in grey, where T is chosen as the median functional to find $\xi(\eta)$, and ι maps η to the smallest index of the elements in $\xi(\eta)$. Note that the displayed epigraph is for a function with its argument on the y-axis

Proposition 2 *Let $\xi : \mathbb{R} \to \mathscr{X}$ be a decreasing, left-continuous function such that $\xi(\eta) \in X(\eta)$, where left-continuity means that if $\eta_n \uparrow \eta$ and $z \in \xi(\eta_n)$, then $z \in \xi(\eta)$. Then, the function $\hat{g} : \mathscr{Z} \to \mathbb{R}$ given by*

$$\inf\{\eta : z \notin \xi(\eta)\} = \hat{g}(z) = \max\{\eta : z \in \xi(\eta)\} \tag{8}$$

is the unique function that satisfies

$$\{z : g(z) \geq \eta\} = \xi(\eta) \quad \text{for all } \eta \in \mathbb{R},$$

among all increasing functions $g : \mathscr{Z} \to \mathbb{R}$.

Proof The left-continuity and monotonicity of $\xi : \mathbb{R} \to \mathscr{X}$ implies the equality of infimum and maximum in equation (8). The monotonicity of $\hat{g}$ follows from the monotonicity of ξ and the fact that ξ takes values being superlevel sets of the partial order on $\mathscr{Z}$. Let $\eta' \in \mathbb{R}$. Then,

$$\text{(i)} \quad \hat{g}(z) \geq \eta' \implies \xi(\hat{g}(z)) \subseteq \xi(\eta') \implies z \in \xi(\eta').$$
$$\text{(ii)} \quad \text{For any } z \in \xi(\eta') : \quad \hat{g}(z) = \max\{\eta : z \in \xi(\eta)\} \geq \eta'.$$

Therefore, $\{z : \hat{g}(z) \geq \eta'\} \subseteq \{z : z \in \xi(\eta')\} \subseteq \{z : \hat{g}(z) \geq \eta'\}$ where the first inclusion follows by (i) and the second by (ii). Uniqueness follows because any hypothetical alternative $\bar{g}$ with $\bar{g}(z') \neq \hat{g}(z')$ for some $z' \in \mathscr{Z}$ leads to the contradiction $\xi(\eta) = \{z : \bar{g}(z) \geq \eta\} \neq \{z : \hat{g}(z) \geq \eta\} = \xi(\eta)$ for all η between $\bar{g}(z')$ and $\hat{g}(z')$. □

As a first result, we characterize minimizers of $s_x(\eta)$ at (7) for a given $\eta \in \mathbb{R}$. The following proposition states necessary and sufficient conditions for the inclusion of an upper set x in the set of minimizing superlevel sets $X(\eta)$. This is the first step towards establishing a link between the level η and the value of the functional T on the corresponding level set, and more elementary, it is also the first step in proving the existence of a decreasing function ξ as specified in Proposition 2.

Proposition 3 *Let $\eta \in \mathbb{R}$. Subject to $x, x' \in \mathscr{X}$, the inclusion $x \in X(\eta)$ holds if and only if*

$$v_{x\setminus x'}(\eta) \leq 0 \quad \text{for all } x' \subsetneq x,$$
$$v_{x'\setminus x}(\eta) \geq 0 \quad \text{for all } x' \supsetneq x.$$

Let $x \in X(\eta)$, $x' \in \mathscr{X}$. If $v_{x\setminus x'}(\eta) = v_{x'\setminus x}(\eta)$, then $x' \in X(\eta)$.

Proof Note that $s_x(\eta) \leq s_{x'}(\eta)$ for all $x' \subsetneq x$ and all $x' \supsetneq x$ holds if and only if $v_{x\setminus x'}(\eta) \leq 0$ for all $x' \subsetneq x$ and $v_{x'\setminus x}(\eta) \geq 0$ for all $x' \supsetneq x$. For the first part of the result, note that $x \in X(\eta)$ implies $s_x(\eta) \leq s_{x'}(\eta)$ for all $x' \subsetneq x$ and all $x' \supsetneq x$. Conversely, let $x \in \mathscr{X}$ be such that the latter condition is satisfied. Then, $s_x(\eta) \leq s_{x'\cap x}(\eta)$ and $s_x(\eta) \leq s_{x'\cup x}(\eta)$ for all $x' \in \mathscr{X}$. By substracting $v_{x\setminus x'}(\eta)$ on both sides of the latter inequality, we have $s_{x\cap x'}(\eta) \leq s_{x'}(\eta)$ for all $x' \in \mathscr{X}$, and hence $x \in X(\eta)$. The second part of the result is immediate after adding $s_{x\cap x'}(\eta)$ to both sides of $v_{x\setminus x'}(\eta) = v_{x'\setminus x}(\eta)$, that is, $s_x(\eta) = s_{x'}(\eta)$. □

The following corollary is of particular importance in the context of total orders, where all admissible superlevel sets are pairwise nested.

Corollary 2 *Let $\eta \in \mathbb{R}$ and $x \in X(\eta)$, $x' \in \mathscr{X}$. If $x' \subsetneq x$ and $v_{x\setminus x'}(\eta) = 0$, then $x' \in X(\eta)$. Analogously, if $x' \supsetneq x$ and $v_{x'\setminus x}(\eta) = 0$, then $x' \in X(\eta)$.*

The next result establishes links between two or more sets of minimizing superlevel sets, that is, between $X(\eta)$ and $X(\eta')$ when $\eta \neq \eta'$. Afterwards, Lemma 3 shows the existence of a decreasing function ξ as specified in Proposition 2.

Lemma 2

(a) *Let $\eta, \eta' \in \mathbb{R}$, $\eta < \eta'$, and $x \in X(\eta)$, $x' \in X(\eta')$. Then, $v_{x'\setminus x}(\eta'') = 0$ for all $\eta'' \in [\eta, \eta']$.*
(b) *Let $\eta \in \mathbb{R}$ and $x', x'' \in X(\eta)$, $x \in \mathscr{X}$. If $x \in \bigcup_{\eta\in\mathbb{R}} X(\eta)$ and $x' \supseteq x \supseteq x''$, then $x \in X(\eta)$.*
(c) *Let $\eta, \eta' \in \mathbb{R}$, $\eta < \eta'$, and $x \in X(\eta)$, $x' \in X(\eta')$. Then, $x \cup x' \in X(\eta)$ and $x \cap x' \in X(\eta')$.*

Proof

(a) We have $(x \cup x') \setminus x = x' \setminus x = x' \setminus (x \cap x')$. The statement is trivial if $x' \setminus x = \emptyset$. Otherwise, $v_{x'\setminus x}(\eta) \geq 0 \geq v_{x'\setminus x}(\eta')$ by Proposition 3, where the statement follows from the monotonicity of the identification function in its first argument.
(b) The statement is trivial if $x = x'$, $x = x''$, or $x \notin X(\eta')$ for all $\eta' \neq \eta$. Therefore, assume $x \in X(\eta')$, $\eta' \neq \eta$. If $\eta < \eta'$, then $v_{x\setminus x''}(\eta) = 0$ by part (a). If $\eta' < \eta$, then $v_{x'\setminus x}(\eta) = 0$ by part (a). In either case, $x \in X(\eta)$ by Corollary 2.

(c) We have $s_x(\eta) \le s_{x\cup x'}(\eta)$ and $s_{x'}(\eta') \le s_{x\cap x'}(\eta')$, and $v_{x'\setminus x}(\eta'') = 0$ for all $\eta'' \in [\eta, \eta']$ by part (a). That means, $s_x(\eta) = s_{x\cup x'}(\eta)$ and $s_{x'}(\eta') = s_{x\cap x'}(\eta')$.

□

Lemma 3

(a) *There exists a decreasing function $\xi : \mathbb{Q} \to \mathscr{X}$ such that $\xi(q) \in X(q)$ for all $q \in \mathbb{Q}$.*
(b) *Let $\eta_n \uparrow \eta$ and $x_n \in X(\eta_n)$, $x_n \supseteq x_{n+1}$. Then, $x = \bigcap_{n\in\mathbb{N}} x_n \in X(\eta)$.*

Proof

(a) Let $\{q_n\} = \mathbb{Q}$ be an enumeration of the rationals. We define $\xi(q_n)$ inductively. Pick $x_1 \in X(q_1)$ and set $\xi(q_1) = x_1$. For $n \ge 2$, define

$$x_n^- = \bigcup_{\substack{i \in \{1,\dots,n-1\} \\ q_i > q_n}} \xi(q_i), \quad x_n^+ = \bigcap_{\substack{i \in \{1,\dots,n-1\} \\ q_i < q_n}} \xi(q_i),$$

if $\{i : q_i > q_n\} \neq \emptyset$ and $\{i : q_i < q_n\} \neq \emptyset$. If $\{i : q_i > q_n\} = \emptyset$, we set $x_n^- = \emptyset$, and if $\{i : q_i < q_n\} = \emptyset$, we set $x_n^+ = \mathscr{X}$. We choose any $x_n \in X(q_n)$ and set $\xi(q_n) = (x_n \cup x_n^-) \cap x_n^+$. At each step n, $\xi(q_n) \in X(q_n)$ follows by 2 (a), and $\xi(q_n) \subseteq x_n^+$. Furthermore, we show by induction that $x_n^- \subseteq \xi(q_n)$ for all n. For $n = 2$, this is easily verified. Suppose the claim holds for $n - 1 \ge 2$. If $q_n > q_{n-1}$, then $x_n^- = x_{n-1}^-$ and $x_n^+ = x_{n-1}^+ \cap \xi(q_{n-1}) = \xi(q_{n-1})$, hence

$$x_n^- = x_{n-1}^- \subseteq \left(x_n \cup x_{n-1}^-\right) \cap \xi\left(q_{n-1}\right) = \xi\left(q_n\right).$$

If $q_n < q_{n-1}$, then $x_n^- = x_{n-1}^- \cup \xi(q_{n-1}) = \xi(q_{n-1})$ and $x_n^+ = x_{n-1}^+$, hence

$$x_n^- = \xi(q_{n-1}) \subseteq (x_n \cup \xi(q_{n-1})) \cap x_{n-1}^+ = \xi(q_n).$$

In summary, for $k < n$, if $q_k < q_n$, then $\xi(q_n) \subseteq x_n^+ \subseteq \xi(q_k)$, and if $q_k > q_n$, $\xi(q_k) \subseteq x_n^- \subseteq \xi(q_n)$ showing that ξ is decreasing.

(b) We have $s_{x_n}(\eta_n) \le s_{x'}(\eta_n)$ for all $x' \in \mathscr{X}$. Furthermore, the definitions of x and V imply $\mathbb{1}\{z \in x_n\}V(\eta_n, y) \to \mathbb{1}\{z \in x\}V(\eta, y)$ pointwise, and we have $\mathbb{1}\{z \in x_n\}V(\eta_n, y) \le \sup_{n\in\mathbb{N}} |V(\eta_n, y)|$. By the dominated convergence theorem, $s_{x_n}(\eta_n) \to s_x(\eta)$ and $s_{x'}(\eta_n) \to s_{x'}(\eta)$.

□

Part (b) of Lemma 3 describes a possible completion step for part (a) that also modifies ξ to be left-continuous. In a nutshell, any decreasing $\xi' : \mathbb{Q} \to \mathscr{X}$ that satisfies $\xi'(\eta') \in X(\eta')$ for all $\eta' \in \mathbb{Q}$ admits a left-continuous version on $\mathbb{R}$, $\xi : \eta \mapsto \bigcap_{\eta'<\eta} \xi'(\eta') \in X(\eta)$, where the intersection is over all $\eta' \in \mathbb{Q}$, $\eta' < \eta$.

In order to prove the existence of a function ξ (and thus $\hat{g}$) that solves the isotonic regression problem, we need that $\mathscr{X}$ is closed under union and intersection. This property is essential for Lemma 3.

We could also start with a set $\mathscr{X}$ of subsets of $\{z_1, \dots, z_n\}$ that are interpreted as the admissible superlevel sets of the function g that is to be fitted. If $\mathscr{X}$ is closed under union and intersection, then $\mathscr{X}$ induces a partial order on $\{z_1, \dots, z_n\}$ by Birkhoff's Representation Theorem; see for example Gurney and Griffin (2011). Consequently, the optimal function $\hat{g}$ always exists and is increasing.

Starting with $\mathscr{X}$, one could formulate constraints other than isotonicity on g as long as they can be formulated in terms of restrictions on admissible superlevel sets. Examples are unimodality with a fixed mode or quasi-convexity with a fixed minimal point. Generally, there is no solution that is simultaneously optimal with respect to all elementary loss functions; see Sect. 3 for an example in the case of a unimodality constraint without a fixed mode.

4.1 Characterization of optimal solutions

The following proposition is essential to provide min-max and max-min bounds on solutions to the isotonic regression problem. We relate the threshold $\eta \in \mathbb{R}$ to the bounds of the functional T on subsets of the data. As a reminder, we write $T^-_x = T^-_{P_x} = \inf T(P_x)$ and $T^+_x = T^+_{P_x} = \sup T(P_x)$.

Proposition 4 *Let $\eta \in \mathbb{R}$, $x \in X(\eta)$. Then, subject to $x' \in \mathscr{X}$,*

$$\max_{x' \supsetneq x} T^-_{x' \setminus x} \leq \eta \leq \min_{x' \subsetneq x} T^+_{x \setminus x'},$$

$$\max_{x' \supsetneq x, x' \notin X(\eta)} T^+_{x' \setminus x} < \eta \leq \min_{x' \subsetneq x, x' \notin X(\eta)} T^-_{x \setminus x'}.$$

Proof For all $x' \supsetneq x$, we have $v_{x' \setminus x}(\eta) \geq 0$. For all $x' \subsetneq x$, we have $v_{x \setminus x'}(\eta) \leq 0$. If $x' \notin X(\eta)$, then both inequalities are strict. Corollary 1 implies the result. □

Figure 5 illustrates the statement in Proposition 4 for a total order in the context of the expectation functional, which is a functional of singleton type. We now state and show one of our main results which is that $\hat{g}$ coincides with or is bounded by a min-max and max-min solution.

Theorem 1 *Let $z \in \mathscr{Z}$ and let $\hat{g}$ be a solution to the isotonic regression problem. Then, subject to $x, x' \in \mathscr{X}$,*

$$\min_{x' : z \notin x'} \max_{x \supsetneq x'} T^-_{x \setminus x'} \leq \hat{g}(z) \leq \max_{x : z \in x} \min_{x' \subsetneq x} T^+_{x \setminus x'}.$$

Proof Applying the first set of bounds from Proposition 4 to the formula for $\hat{g}$ at (8), we obtain

$$\inf_{\eta : z \notin \xi(\eta)} \max_{x \supsetneq \xi(\eta)} T^-_{x \setminus \xi(\eta)} \leq \hat{g}(z) \leq \max_{\eta : z \in \xi(\eta)} \min_{x' \subsetneq \xi(\eta)} T^+_{\xi(\eta) \setminus x'}.$$

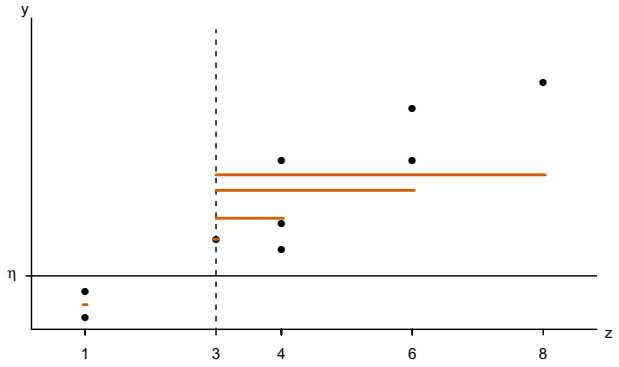

Fig. 5 Minimizing indices are separators. For a sample of 9 data points, the graph illustrates the functional value (expectation) on relevant subsets of the data for a given η and the minimizing index $i = 3$. The expectation value (vertical location of a brown line) is above or below η when the corresponding subsample extends (horizontal extension of a brown line) to the right or left of the minimizing index, respectively

The lower bound is bounded from below by $\min_{x':z\notin x'} \max_{x\supsetneq x'} T^-_{x\setminus x'}$, and the upper bound is bounded from above by $\max_{x:z\in x} \min_{x'\subsetneq x} T^+_{x\setminus x'}$. □

The previous statement is closely related to the coinciding max-min and min-max solutions at (2) for the expectation functional and a total order isotonicity constraint. For an analogous statement of uniqueness, as referred to in Example 2, we need the following lemma on a modified max-min inequality in the context of partial orders.

Lemma 4 *Suppose that T is of singleton type. Let $z \in \mathscr{Z}$ be such that $P(\{z\} \times \mathbb{R}) > 0$. Then, subject to $x, x' \in \mathscr{X}$,*

$$\max_{x:z\in x} \min_{x'\subsetneq x} T^+_{x\setminus x'} \leq \min_{x':z\notin x'} \max_{x\supsetneq x'} T^-_{x\setminus x'}.$$

Proof Let $x'' \in \mathscr{X}$ such that $z \notin x''$, then

$$\max_{x:z\in x} \min_{x'\subsetneq x} T^+_{x\setminus x'} = \max_{x:z\in x} \min_{\substack{x'\subsetneq x \\ P((x\setminus x')\times\mathbb{R})>0}} T^+_{x\setminus x'} = \max_{x:z\in x} \min_{\substack{x'\subsetneq x \\ P((x\setminus x')\times\mathbb{R})>0}} T^-_{x\setminus x'}$$
$$\leq \max_{x:z\in x} T^-_{x\setminus(x\cap x'')} = \max_{x:z\in x} T^-_{(x\cup x'')\setminus x''} \leq \max_{x:x\supsetneq x''} T^-_{x\setminus x''},$$

where the last inequality holds because $x \cup x'' \in \mathscr{X}$ and if $z \in x$ then $x \cup x'' \supsetneq x''$. □

In general, a similar statement on coinciding max-min and min-max solutions always holds, where the choice of ξ determines whether $\hat{g}$ attains the minimal or maximal elements of the functional. It is possible to define minimal and maximal solutions. Recall that we defined $X(\eta)$ as the set of superlevel sets $x \in \mathscr{X}$ minimizing $s_x(\eta)$ at (7). Let

$$X^-(\eta) = \{x \in X(\eta) : \nexists\, x' \in X(\eta) \text{ such that } x' \subsetneq x\},$$
$$X^+(\eta) = \{x \in X(\eta) : \nexists\, x' \in X(\eta) \text{ such that } x' \supsetneq x\}$$

denote the sets of minimal and maximal elements of $X(\eta)$, respectively.

Proposition 5 *Let $z \in \mathscr{Z}$ be such that $P(\{z\} \times \mathbb{R}) > 0$, and let $\xi : \mathbb{R} \to \mathscr{X}$ be decreasing and left-continuous.*

(a) *If $\xi(\eta) \in X^+(\eta)$ for all $\eta \in \mathbb{R}$, then, subject to $x, x' \in \mathscr{X}$,*

$$\hat{g}(z) = \min_{x' : z \notin x'} \max_{x \supsetneq x'} T^+_{x \setminus x'} = \max_{x : z \in x} \min_{x' \subsetneq x} T^+_{x \setminus x'}.$$

(b) *If $\xi(\eta) \in X^-(\eta)$ for all $\eta \in \mathbb{R}$, then, subject to $x, x' \in \mathscr{X}$,*

$$\hat{g}(z) = \min_{x' : z \notin x'} \max_{x \supsetneq x'} T^-_{x \setminus x'} = \max_{x : z \in x} \min_{x' \subsetneq x} T^-_{x \setminus x'}.$$

Proof The proof follows using Lemma 4 and applying the same steps as in the proof of Theorem 1 to the second set of bounds in Proposition 4. □

Let us denote the solution in part (a) of Proposition 5 by g^+ and the one in part (b) by g^-. Clearly, it always holds that $g^- \le g^+$. It is a natural question whether any increasing function g that satisfies $g^- \le g \le g^+$ is also a minimizer of the criterion (6). It turns out that the answer is negative; see Mösching and Dümbgen (2020, Remark 2.2, Example 2.4). Combining Propositions 5 to 8 and Corollary 3, gives a complete characterizations of all possible solutions to the isotonic regression problem for partial orders. For the following results, it is not required that g^-, g^+ are the solutions from Proposition 5. Unless specified, they do not even need to satisfy $g^- \le g^+$ everywhere. We define $\xi^- : \eta \mapsto \{z : g^-(z) \ge \eta\}$ and ξ^+ analogously.

Proposition 6 *Let g^- and g^+ be two solutions to the isotonic regression problem such that $g^- \le g^+$. Let $\hat{g}$ be isotonic, $g^- \le \hat{g} \le g^+$, and suppose that all superlevel sets of $\hat{g}$ lie in $\bigcup_{\eta \in \mathbb{R}} X(\eta)$. Then, $\hat{g}$ is a solution to the isotonic regression problem.*

Proof For $\eta \in \mathbb{R}$ define $\xi(\eta) = \{z : \hat{g}(z) \ge \eta\}$. The functions ξ, ξ^-, ξ^+ are decreasing, that is $\xi(\eta) \supseteq \xi(\eta')$ for $\eta \le \eta'$, and left-continuous. For ξ^-, ξ^+ it holds that $\xi^-(\eta)$, $\xi^+(\eta) \in X(\eta)$. Since, for all $z \in \mathscr{Z}$, it holds that

$$\begin{aligned} g^-(z) = \max\{\eta : z \in \xi^-(\eta)\} \le g(z) = \max\{\eta : z \in \xi(\eta)\} \\ \le g^+(z) = \max\left\{\eta : z \in \xi^+(\eta)\right\}, \end{aligned}$$

we obtain $\xi^-(\eta) \subseteq \xi(\eta) \subseteq \xi^+(\eta)$ for all $\eta \in \mathbb{R}$. Lemma 2 (b) implies the result. □

The following corollary is an immediate consequence of Lemma 2 (c).

Corollary 3 *Let g^- and g^+ be two solutions to the isotonic regression problem. Then, the distributive lattice generated by ξ^- and ξ^+ is a subset of $\bigcup_{\eta\in\mathbb{R}} X(\eta)$.*

Having two solutions g^- and g^+ allows us to find all solutions to the isotonic regression problem with superlevel sets that lie in the lattice generated by ξ^- and ξ^+. Examples include solutions that transition from g^- to g^+ at a particular threshold η,

$$\hat{g}(z) = \begin{cases} g^+(z), & z \in \xi^+(\eta), \\ g^-(z), & \text{otherwise}, \end{cases}$$

or pointwise convex combinations of solutions with $\alpha \in (0, 1)$,

$$\hat{g}(z) = \alpha g^-(z) + (1-\alpha) g^+(z).$$

In order to refine the lattice of minimizing upper sets from Corollary 3 with the purpose to characterize all solutions, we pose the question whether simple separation rules exist for the set difference of consecutive lattice elements. These sets necessarily take the form of the intersection of a level set of g^- and a level set of g^+, that is, sets of the form $\{z : g^-(z) = \eta^- \text{ and } g^+(z) = \eta^+\}$. These rules do exist as we show in Propositions 7 and 8. First, we introduce the notion of a separation.

Definition 3 A *separation* of a set $Z \in \mathscr{P}(\mathscr{Z})$ is a collection of sets $Z_1, \ldots, Z_n \subseteq Z$ that are pairwise separated and satisfy $Z = \bigcup_{i=1}^n Z_i$. Two sets Z_i and Z_j are *separated* if for all $z' \in Z_i$ and $z'' \in Z_j$, neither $z' \preceq z''$ nor $z'' \preceq z'$.

Proposition 7 *Let g^- and g^+ be two solutions to the isotonic regression problem, and let $\eta^-, \eta^+ \in \mathbb{R}$, $\eta^- < \eta^+$, be such that $Z = \{z : g^-(z) = \eta^- \text{ and } g^+(z) = \eta^+\}$ is nonempty. Furthermore, let $Z_1, \ldots, Z_n$ be a separation of Z, and let $x' = \xi^-(\eta^-) \cap \xi^+(\eta^+)$ and $x'' = x' \setminus Z$. Then, $x'' \cup Z_k \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$, $k = 1, \ldots, n$.*

Proof Without loss of generality, we show the claim for $k = 1$. By Lemma 2 (c), we have $x' \in X(\eta^+)$ and $x'' = \xi^-(\eta^- + \epsilon_1) \cup \xi^+(\eta^+ + \epsilon_2) \in X(\eta^- + \epsilon_1)$ for some $\epsilon_1, \epsilon_2 > 0$. More precisely, we have $x', x'' \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ by Lemma 2 (b), since $\xi^-(\eta) \subseteq x'' \subseteq x' \subseteq \xi^+(\eta)$, $\eta \in (\eta^-, \eta^+]$.

Let $x_1 = x'' \cup Z_1$ and $x_2 = x' \setminus Z_1$ both of which are upper sets in $\mathscr{X}$. Then $Z_1 = x_1 \setminus x''$ but also $Z_1 = x' \setminus x_2$. Therefore, $v_{Z_1}(\eta) \geq 0 \geq v_{Z_1}(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ by Proposition 3. Then the statement follows from Corollary 2. □

Proposition 7 allows us to find additional solutions to the isotonic regression problem with superlevel sets where separation elements have been added to known minimizing superlevel sets. Using the variables defined in Proposition 7, one example of a new solution is

$$\hat{g}(z) = \begin{cases} \eta, & z \in Z_1, \\ g^+(z), & z \in x'', \\ g^-(z), & \text{otherwise}, \end{cases}$$

where $\eta \in (\eta^-, \eta^+]$. Iterative application of Proposition 7 recovers all minimizing superlevel sets that can be obtained from the solutions in Proposition 5 via Corollary 3 and the information on the partially ordered set $\mathscr{X}$.

Proposition 8 allows us to recover the remaining minimizing superlevel sets when the distribution P of the random vector (Z, Y) is fully known. Again, P may be the empirical distribution for a series of covariate-response pairs. In fact, this proposition is a generalization of Proposition 7 that determines whether a level set intersection of g^- and g^+ can be split further by calculating values of the lower bound of the functional T.

Proposition 8 *Let g^- and g^+ be two solutions to the isotonic regression problem, and let $\eta^-, \eta^+ \in \mathbb{R}$, $\eta^- < \eta^+$, be such that $Z = \{z : g^-(z) = \eta^- \text{ and } g^+(z) = \eta^+\}$ is non-empty. Furthermore, let $x' = \xi^-(\eta^-) \cap \xi^+(\eta^+)$ and $x'' = x' \setminus Z$. For $x \in \mathscr{X}$, $x' \supsetneq x \supsetneq x''$, we have $T^-_{x'\setminus x} \leq \eta^-$ if and only if $x \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$.*

Proof We have $x', x'' \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ as in the proof of Proposition 7. Then, $v_{x'\setminus k}(\eta^+) \leq 0$ for all $k \in \mathscr{X}$, $k \subsetneq x'$, by Proposition 3, and hence $T^+_{x'\setminus k} \geq \eta^+$ by Corollary 1. Analogously, $v_{k\setminus x''}(\eta) \geq 0$ for all $k \in \mathscr{X}$, $k \supsetneq x''$, $\eta \in (\eta^-, \eta^+]$, leading to $T^-_{k\setminus x''} \leq \eta^-$.

For the first part of the statement, let $x \in \mathscr{X}$, $x' \supsetneq x \supsetneq x''$, be such that $T^-_{x'\setminus x} \leq \eta^-$. We show that $x \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ using Proposition 3. We have $T^+_{x'\setminus k} \leq \max\{T^-_{x'\setminus x}, T^+_{x\setminus k}\}$ for all $k \subsetneq x$ by Lemma 1. Since $T^-_{x'\setminus x} \leq \eta^-$ by assumption and as just shown $T^+_{x'\setminus k} \geq \eta^+$, we obtain $T^+_{x\setminus k} \geq \eta^+$. By Corollary 1, $v_{x\setminus k}(\eta) \leq 0$ for all $k \subsetneq x$, $\eta \leq \eta^+$, that is, the first inequality in Proposition 3 holds for all $\eta \in (\eta^-, \eta^+]$. Similarly, $T^-_{k\setminus x''} \geq \min\{T^-_{k\setminus x}, T^+_{x\setminus x''}\}$ for all $k \supsetneq x$. Since $T^-_{k\setminus x''} \leq \eta^-$ and $T^+_{x\setminus x''} \geq \eta^+$, we obtain $T^-_{k\setminus x} \leq \eta^-$. Therefore, $v_{k\setminus x}(\eta) \geq 0$, for all $\eta > \eta^-$, $k \supsetneq x$, that is, the second inequality in Proposition 3 holds for all $\eta \in (\eta^-, \eta^+]$.

To prove the converse, note that $x \in X(\eta)$ for all $\eta \in (\eta^-, \eta^+]$ implies that $v_{k\setminus x}(\eta) \geq 0$ for all $\eta \in (\eta^-, \eta^+]$, $k \supsetneq x$. Hence, in particular, $v_{x'\setminus x}(\eta) \geq 0$ and $T^-_{x'\setminus x} \leq \eta$ for all $\eta \in (\eta^-, \eta^+]$, and, therefore $T^-_{x'\setminus x} \leq \eta^-$. □

4.2 Pool-adjacent-violators algorithm

This section discusses the PAV algorithm as adapted to the context of set-valued functionals and shows the optimality of its solution using the methods introduced in this paper. The algorithm solves the isotonic regression problem for a total order, taking observations $(z_1, y_1), \ldots, (z_n, y_n)$, $z_1 < \cdots < z_n$. In general, the PAV

algorithm only applies to totally ordered covariates but we comment on extensions to partial orders at the end of this section.

Algorithm 1: Lazy PAV for set-valued functionals

Result: A partition $\mathcal{Q}_{\mathrm{PAV}}$
Initialize $\mathcal{Q} = \{Q_1, \dots, Q_n\}$ where $Q_i = \{z_i\}$, endow $\mathcal{Q}$ with the order induced by $z_1 < \cdots < z_n$, and begin the iteration with $Q \leftarrow Q_n$;
loop
  while Q *has a predecessor* $\mathrm{pred}(Q)$ *in* $\mathcal{Q}$ ***and*** $T^-_{\mathrm{pred}(Q)} > T^+_Q$ **do**
    merge its predecessor into Q, and update the partition $\mathcal{Q}$;
  while Q *has a successor* $\mathrm{S}(Q)$ *in* $\mathcal{Q}$ ***and*** $T^-_Q > T^+_{\mathrm{S}(Q)}$ **do**
    merge its successor into Q, and update the partition $\mathcal{Q}$;
  if Q *has a predecessor in* $\mathcal{Q}$ **then** $Q \leftarrow \mathrm{pred}(Q)$ **else** end loop;

We describe a lazy version of the PAV algorithm for set-valued functionals in Algorithm 1. It is lazy in the sense that it only creates a partition $\mathcal{Q}_{\mathrm{PAV}}$ of $\{z_1, \dots, z_n\}$ without returning the isotonic solution, and it stops as soon as an increasing function $g : \{z_1, \dots, z_n\} \to \mathbb{R}$ exists that is constant on each element of the current partition $\mathcal{Q}$ and satisfies

$$g(z) \in T(P_Q) \quad \text{for all } Q \in \mathcal{Q} \text{ and } z \in Q, \tag{9}$$

that is, when no further pooling is necessary. The solution g_{PAV} that satisfies the previous requirements is unique for functionals of singleton type, essentially given by (9) for the partition $\mathcal{Q}_{\mathrm{PAV}}$. When choosing a solution for functionals that are of interval type, additional steps are required that ensure monotonicity because neither upper nor lower bounds of the functional intervals are necessarily nondecreasing on $\mathcal{Q}_{\mathrm{PAV}}$. For the sake of brevity, we assume that a valid solution g_{PAV} has been chosen.

To show the optimality of the solution of the PAV algorithm, the first and most apparent property that we observe is that for all $z \in \{z_1, \dots, z_n\}$, $Q_1, Q_2 \in \mathcal{Q}_{\mathrm{PAV}}$, $\min Q_1 \le z \le \max Q_2$, we have

$$T^-_{Q_1} \le g_{\mathrm{PAV}}(z) \le T^+_{Q_2}, \tag{10}$$

since otherwise either g_{PAV} is not increasing or the condition (9) is violated. Definition 1 and its Corollary 1 allow for an immediate proof of an additional property of $\mathcal{Q}_{\mathrm{PAV}}$.

Proposition 9 *Let $\mathcal{Q}$ be a partition of $\{z_1, \dots, z_n\}$ found by the PAV algorithm, $Q \in \mathcal{Q}$, and $z \in Q$. Then,*

$$T^-_{Q|_{\ge z}} \le T^-_Q \le T^+_Q \le T^+_{Q|_{\le z}},$$

where $Q|_{\ge z}$ and $Q|_{\le z}$ denote the restrictions to the elements $q \in Q$ satisfying $q \ge z$ and $q \le z$, respectively.

Proof The second inequality is trivial. For the first inequality, suppose the contrary: There exist $\eta \in \mathbb{R}$, $z \in Q$ such that $T_Q^- < \eta < T_{Q|_{\geq z}}^-$. This implies that $Q \neq Q|_{\geq z}$ and $v_Q(\eta) \geq 0 > v_{Q|_{\geq z}}(\eta)$, hence $v_{Q|_{<z}}(\eta) > 0$. Therefore, $T_{Q|_{<z}}^+ < \eta < T_{Q|_{\geq z}}^-$, which means that Q can be seen as the result of an invalid pooling of $Q|_{<z}$ and $Q|_{\geq z}$. A similar argument applies to the third inequality. □

To show the connection between a valid solution by the PAV algorithm and the score optimizing solution $\hat{g}$ in Sect. 4, we define

$$\xi_{\text{PAV}}(\eta) = \left\{z : \eta \leq g_{\text{PAV}}(z)\right\}, \tag{11}$$

which are necessarily sets of the form $\{z_i, \ldots, z_n\}$. Plugging ξ_{PAV} into the definition of $\hat{g}$ recovers g_{PAV},

$$\begin{aligned}\hat{g}(z) &= \max\left\{\eta : z \in \xi_{\text{PAV}}(\eta)\right\} \\ &= \max\left\{\eta : \eta \leq g_{\text{PAV}}(z)\right\} = g_{\text{PAV}}(z).\end{aligned}$$

In order to show that g_{PAV} solves the isotonic regression problem, it remains to be shown that $\xi_{\text{PAV}}(\eta) \in X(\eta)$ for all $\eta \in \mathbb{R}$.

Proposition 10 *Let $\eta \in \mathbb{R}$, then $\xi_{\text{PAV}}(\eta) \in X(\eta)$.*

Proof Let $\eta \in \mathbb{R}$ and $x = \xi_{\text{PAV}}(\eta)$. For all $Q \in \mathscr{Q}_{\text{PAV}}$, we have $T_{Q\setminus x}^- < \eta \leq T_{Q\cap x}^+$ by statement (10) and defining equality (11). Recall that $T_\emptyset^-$ and $T_\emptyset^+$ are $-\infty$ and ∞, respectively. We now use that $T_{P_1+P_2}^- \leq \max\{T_{P_1}^-, T_{P_2}^-\}$ and $T_{P_1+P_2}^+ \geq \min\{T_{P_1}^+, T_{P_2}^+\}$ for nonnegative measures P_1 and P_2 on $\mathbb{R}$, which is an immediate consequence of Definition 1. Together with Proposition 9, and subject to x' denoting an upper set of the form $\{z_i, \ldots z_n\}$, we have $T_{x'\setminus x}^- \leq \max_{Q\in\mathscr{Q}_{\text{PAV}}} T_{Q\setminus x}^-$ for all $x' \supsetneq x$, and $T_{x\setminus x'}^+ \leq \min_{Q\in\mathscr{Q}_{\text{PAV}}} T_{Q\cap x}^+$ for all $x' \subsetneq x$. Therefore, $v_{x'\setminus x}(\eta) \geq 0$ for all $x' \supsetneq x$, and $v_{x\setminus x'}(\eta) \leq 0$ for all $x' \subsetneq x$, and the statement follows from Proposition 3. □

As a side note, we point out that ξ_{PAV} corresponds to coarsest partition that allows the solution g_{PAV}. Any elements of the partition $\mathscr{Q}_{\text{PAV}}$ from Algorithm 1 on which g_{PAV} takes the same value have been pooled.

Finally, there is the question of a PAV algorithm for general partial orders, where the computational performance is a major problem because no simple rule is known that describes the order in which to resolve violations. In contrast, for a total order we can resolve any violation at a given iteration of the algorithm. However, the most common implementations perform a single pass from front to back, or back to front as in Algorithm 1. The direction becomes relevant once considering a partial order that can be represented as a directed tree. As demonstrated by Thompson (1962) and Pardalos and Xue (1999) violations should be resolved starting from the leaves and at each branching the most severe violation should be resolved first. All orders used in Example 2 and for Fig. 3 can be represented as directed trees. For general partial orders, the most recent version of a

PAV algorithm seems to have been proposed by Sysoev et al. (2011), a segmentation-based approach reconciling a set of local approximate solutions to yield a highly accurate solution that is not guaranteed to be exact.

4.3 Partitioning the covariate set

In Sect. 4.2, we discussed how the PAV algorithm creates a partition of $\mathscr{X}$, and that it leads to a solution $\hat{g}$ of the isotonic regression problem in the context of total orders. In this section, we show how a solution to the isotonic regression problem leads to a corresponding partition $\mathscr{Q}$ of $\mathscr{X}$, such that the solution satisfies

$$\hat{g}(z) \in T(P_Q), \quad \text{for all } Q \in \mathscr{Q},\ z \in Q,$$

and the solution is constant on every element of the partition. Let T be a functional of singleton type, and $\hat{g}$ be a solution to the isotonic regression problem. Subject to $x, x', k, k' \in \mathscr{X}$, the combination of Theorem 1 and Lemma 4 yields

$$\begin{aligned}\hat{g}(z) &= \max_{x: z \in x} \min_{x' \subsetneq x} T^{+}_{x \setminus x'} \\ &= \min_{k': z \notin k'} \max_{k \supsetneq k'} T^{-}_{k \setminus k'}\end{aligned}$$

for all $z \in \mathscr{X}$ with $P(\{z\} \times \mathbb{R}) > 0$. We call (x, x') a max-min pair for z if $z \in x$, $x' \subsetneq x$, and $\hat{g}(z) = T^{+}_{x \setminus x'}$, and we call (k', k) a min-max pair for z if $z \notin k'$, $k \supsetneq k'$, and $\hat{g}(z) = T^{-}_{k \setminus k'}$. For a pair $x, x' \in \mathscr{X}$ such that $T^{-}_{x \setminus x'} = T^{+}_{x \setminus x'}$, we also use the notation $T^{\pm}_{x \setminus x'}$. Note that for a functional T of singleton type, we have $T(P_{x \setminus x'}) = \{T^{\pm}_{x \setminus x'}\}$ if $P((x \setminus x') \times \mathbb{R}) > 0$. The following lemma provides the necessary tools to construct the partition $\mathscr{Q}$.

Lemma 5 *Let T be a functional of singleton type, and $\hat{g}$ be a solution to the isotonic regression problem. Furthermore, let $z \in \mathscr{X}$ such that $P(\{z\} \times \mathbb{R}) > 0$, and let $(x_1, x_1'), (x_2, x_2')$ be max-min pairs for z, and $(k_1', k_1), (k_2', k_2)$ be min-max pairs for z. Then the following statements hold*:

(a) *We have that* $\hat{g}(z) = T^{\pm}_{x_1 \setminus k_1'} = T^{\pm}_{(x_1 \cup x_2) \setminus k_1'} = T^{\pm}_{x_1 \setminus (k_1' \cap k_2')}$.
(b) *If $x, k' \in \mathscr{X}$ such that $z \in x$, $z \notin k'$, and $\hat{g}(z) = T^{\pm}_{x \setminus k'}$, then $(x, x \cap k')$ is a max-min pair for z, and $(k', k' \cup x)$ is a min-max pair for z.*
(c) *If $\tilde{z} \in x_1 \setminus k_1'$, then (x_1, x_1') is a max-min pair for $\tilde{z}$, and (k_1', k_1) is a min-max pair for $\tilde{z}$.*

Proof We repeatedly use the inequalities $\hat{g}(z) = T^{+}_{x_1 \setminus x_1'} = \min_{x' \in \mathscr{X}} T^{+}_{x_1 \setminus x'} \le T^{+}_{x_1 \setminus k'}$ and $\hat{g}(z) = T^{-}_{k_1 \setminus k_1'} = \max_{k \in \mathscr{X}} T^{-}_{k \setminus k_1'} \ge T^{-}_{x \setminus k_1'}$ for all $x, k' \in \mathscr{X}$, where the second equality holds because $T^{+}_{P} = \infty$ and $T^{-}_{P} = -\infty$ for null measures P. Furthermore, by

assumption, $T(P_{x\setminus k'})$ is a singleton if $P((x \setminus k') \times \mathbb{R}) > 0$, and therefore, $T(P_{x\setminus k'})$ is a singleton if $z \in x$ and $z \notin k'$.

(a) Clearly, $z \in x_1$, $z \in x_2$, $z \notin k_1'$, and $z \notin k_2'$. Hence, $\hat{g}(z) \leq T^{\pm}_{x_1\setminus k_1'} \leq \hat{g}(z)$ implies the first statement. Furthermore, $\hat{g}(z) \leq T^{+}_{x_2\setminus(x_1\cup k_1')} = T^{+}_{(x_2\setminus x_1)\setminus k_1'}$, and hence $\hat{g}(z) = \min\{T^{-}_{x_1\setminus k_1'}, T^{+}_{(x_2\setminus x_1)\setminus k_1'}\} \leq T^{\pm}_{(x_1\cup x_2)\setminus k_1'} \leq \hat{g}(z)$ confirms the second statement using Lemma 1. Similarly, for the third statement, $\hat{g}(z) \leq T^{\pm}_{x_1\setminus(k_1'\cap k_2')} \leq \max\{T^{+}_{x_1\setminus k_1'}, T^{-}_{(x_1\cap k_1')\setminus k_2'}\} = \hat{g}(z)$.

(b) The statement follows immediately from $T^{-}_{x\setminus k'} = T^{+}_{x\setminus k'}$, $(x \cup k') \setminus k' = x \setminus k' = x \setminus (x \cap k')$, and the definition of max-min and min-max pairs.

(c) Let $(x_{\tilde{z}}, x_{\tilde{z}}')$ be a max-min pair for $\tilde{z}$ and $(k_{\tilde{z}}', k_{\tilde{z}})$ be a min-max pair for $\tilde{z}$. Then the statement follows from $\hat{g}(z) \leq T^{\pm}_{x_1\setminus k_{\tilde{z}}'} \leq \hat{g}(\tilde{z}) \leq T^{\pm}_{x_{\tilde{z}}\setminus k_1'} \leq \hat{g}(z)$.

□

Proposition 11 *Let T be a functional of singleton type. Then there exists a partition $\mathscr{Q}$ of $\mathscr{Z}$ such that $\hat{g}$ is constant on every element of the partition almost everywhere and $\hat{g}(z) \in T(P_Q)$ for all $Q \in \mathscr{Q}$, $z \in Q$ such that $P(\{z\} \times \mathbb{R}) > 0$.*

Proof Let $\bar{x}_z$ denote the union of the first components of all max-min pairs for $z \in \mathscr{Z}$, and let $\bar{k}_z'$ denote the intersection of the first components of all min-max pairs for $z \in \mathscr{Z}$. By Lemma 5 (a), we have $\hat{g}(z) = T^{\pm}_{\bar{x}_z\setminus\bar{k}_z'}$. We now show that the collection $\mathscr{Q}$ of sets $Q_z = \bar{x}_z \setminus \bar{k}_z'$ is a partition of $\mathscr{Z}$. First, we have $\bigcup_{z\in\mathscr{Z}} Q_z = \mathscr{Z}$, since $z \in \bar{x}_z$ and $z \notin \bar{k}_z'$ for all $z \in \mathscr{Z}$. Second, by Lemma 5 (b), we have that $(\bar{x}_z, \bar{x}_z \cap \bar{k}_z')$ is a max-min pair for z and $(\bar{k}_z', \bar{k}_z' \cup \bar{x}_z)$ is a min-max pair for z. Then, by Lemma 5 (c), we have $\bar{x}_z \subset \bar{x}_{\tilde{z}}$ and $\bar{k}_z' \supset \bar{k}_{\tilde{z}}'$ for all $\tilde{z} \in Q_z$, i.e., $Q_z \subset Q_{\tilde{z}}$ and in particular $z \in Q_{\tilde{z}}$. Swapping the roles of z and $\tilde{z}$ gives $Q_{\tilde{z}} \subset Q_z$. Therefore, $Q_z = Q_{\tilde{z}}$ for all $z \in \mathscr{Z}, \tilde{z} \in Q_z$. □

When T is a functional of interval type, we therefore obtain a partition for every fixed convex combination of its lower bound T^- and its upper bound T^+.

Acknowledgements We would like to thank two reviewers, Tilmann Gneiting, Alexandre Mösching and Lutz Dümbgen for inspiring discussions and valuable comments. Alexander I. Jordan acknowledges the support of the Klaus Tschira Foundation. Anja Mühlemann and Johanna F. Ziegel gratefully acknowledge financial support from the Swiss National Science Foundation.

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics, 26,* 641–647.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., Brunk, H. D. (1972). *Statistical inference under order restrictions*. London: Wiley.

Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika, 46,* 36–48.

Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives. *II. Biometrika, 46,* 328–335.

Bellec, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics, 46,* 745–780.

Brümmer, N., Du Preez, J. (2013). The PAV algorithm optimizes binary proper scoring rules. arXiv:1304.2331.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Annals of Mathematical Statistics, 26,* 607–616.

Dawid, A. P. (2016). Contribution to the discussion of Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings by Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. *Journal of the Royal Statistical Society. Series B. Statistical Methodology, 78,* 505–562.

Ehm, W., Gneiting, T., Jordan, A., Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society. Series B. Statistical Methodology, 78,* 505–562.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association, 106,* 746–762.

Groeneboom, P., Jongbloed, G. (2014). *Nonparametric estimation under shape constraints*. New York: Cambridge University Press.

Guntuboyina, A., Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science, 33*, 568–594.

Gurney, A. J. T., Griffin, T. G. (2011). Pathfinding through congruences. *Relational and Algebraic Methods in Computer Science* (Vol. 6663, pp. 180–195). Heidelberg: Springer.

Han, Q., Wang, T., Chatterjee, S., Samworth, R. J. (2019). Isotonic regression in general dimensions. *The Annals of Statistics, 47*, 2440–2471.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*, 73–101.

Kyng, R., Rao, A., Sachdeva, S. (2015). Fast, provable algorithms for isotonic regression in all L_p-norms. *Advances in Neural Information Processing Systems 28* (pp. 2719–2727). Red Hook: Curran Associates Inc.

Luss, R., Rosset, S. (2014). Generalized isotonic regression. *Journal of Computational and Graphical Statistics, 23*, 192–210.

Luss, R., Rosset, S. (2017). Bounded isotonic regression. *Electronic Journal of Statistics, 11,* 4488–4514.

Miles, R. E. (1959). The complete amalgamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika, 46*, 317–327.

Mösching, A., Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics, 14*, 24–49.

Newey, W. K., Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica, 55*, 819–847.

Pardalos, P. M., Xue, G. (1999). Algorithms for a class of isotonic regression problems. *Algorithmica, 23*, 211–222.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics, 160*, 246–256.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics, 38,* 796–809.

Polonik, W. (1998). The silhouette, concentration functions and ML-density estimation under order restrictions. *The Annals of Statistics, 26*, 1857–1877.

Robertson, T., Wright, F. T. (1973). Multiple isotonic median regression. *The Annals of Statistics, 1*, 422–432.

Robertson, T., Wright, F. T. (1980). Algorithms in order restricted statistical inference and the Cauchy mean value property. *The Annals of Statistics, 8*, 645–651.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association, 66*, 783–801.

Stout, Q. F. (2015). Isotonic regression for multiple independent variables. *Algorithmica, 71*, 450–470.

Sysoev, O., Burdakov, O., Grimvall, A. (2011). A segmentation-based algorithm for large-scale partially ordered monotonic regression. *Computational Statistics & Data Analysis, 55*, 2463–2476.

Thompson, W. A., Jr. (1962). The problem of negative estimates of variance components. *Annals of Mathematical Statistics, 33,* 273–289.

van Eeden, C. (1958). *Testing and estimating ordered parameters of probability distributions*. Amsterdam: Mathematical Centre.

Ziegel, J. F. (2016). Contribution to the discussion of Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings by Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. *Journal of the Royal Statistical Society. Series B. Statistical Methodology, 78,* 505–562.