Electronic Journal of Statistics
Vol. 16 (2022) 3836–3860
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2034

Isotonic regression for elicitable functionals and their Bayes risk*

Anja Mühlemann and Johanna Ziegel

Institute of Mathematical Statistics and Actuarial Science,
University of Bern, Switzerland
e-mail: anja.muehlemann@stat.unibe.ch; johanna.ziegel@stat.unibe.ch

Abstract: We study the non-parametric isotonic regression problem for bivariate elicitable functionals that are given as an elicitable univariate functional and its Bayes risk. Prominent examples for functionals of this type are (mean, variance) and (Value-at-Risk, Expected Shortfall), where the latter pair consists of important risk measures in finance. We present our results for totally ordered covariates but extenstions to partial orders are given in the appendix.

Received June 2021.

1. Introduction

In isotonic regression, the aim is to fit an increasing function g_1 to observations $(z_1, y_1), \dots, (z_n, y_n)$ such that a chosen loss function is minimized by g_1. The covariates $z_1, \dots, z_n$ take values in some partially ordered space, that is, they can be univariate, multivariate, or even more general. The solution g_1 is then called an optimal solution to the isotonic regression problem. If g_1 is supposed to model a conditional mean, then the loss function should be consistent for the mean in the sense of Gneiting (2011, Definition 1) with a prominent example being the squared error loss. More generally, if g_1 is a model for a conditional functional T, then the loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ should be chosen consistent for this functional T, that is, $\mathbb{E}_P L(t, Y) \le \mathbb{E}_P L(x, Y)$ for all relevant probability distributions P, all $t \in T(P)$ and all $x \in \mathbb{R}$. Loss L is called strictly consistent if the above inequality is strict for all $x \notin T(P)$. This notion of consistency is a property of the functional T and the loss function L and should not to be confused with consistency of an estimator. Strict consistency of L ensures that a correctly specified model minimizes the expected loss at the population level.

If a functional T, that is, a map on a certain class of probability distributions, has a strictly consistent loss function, it is called elicitable. We say that the loss function elicits T. Elicitability is important for forecast comparison (Gneiting, 2011), and yields natural estimation procedures. Unfortunately, some ubiquitous functionals are not elicitable with prominent examples given by the variance (var) and expected shortfall (ES_α), the latter being an important risk measure in finance and insurance. However, although ES_α is not elicitable, it is jointly elicitable together with the α-quantile (q_α); see Fissler and Ziegel (2016) and

*The authors gratefully acknowledge funding from the Swiss National Science Foundation.

Example 2.2. Similarly, while var itself is not elicitable, it is jointly elicitable with the mean ($\mathbb{E}$). This means that both ES_α and var are 2-elicitable, that is, they can both be obtained as a function of a 2-dimensional elicitable functional. In a nutshell, the *elicitation complexity* of a functional is the minimal number k of dimensions needed for the functional to be k-elicitable. Since both ES_α and var are not elicitable themselves but they are 2-elicitable, their elicitation complexity equals 2 (Frongillo and Kash, 2020, Corollary 1 and 3).

Isotonic regression for one-dimensional elicitable functionals is well-understood (Barlow et al., 1972). An interesting aspect is its robustness with respect to the choice of the consistent loss function in the minimization problem. In other words, no matter which strictly consistent loss function we choose for the functional T, we will obtain the same isotonic solution (Brümmer and Du Preez, 2013; Jordan, Mühlemann and Ziegel, 2020). This is in stark contrast to estimation in parametric regression models. In finite samples or for misspecified models, the choice of the consistent loss function may lead to miscellaneous estimates (Patton, 2020).

In this article, we investigate non-parametric regression for bivariate functionals $\underline{T}$ under isotonicity constraints. In particular, we show that simultaneous optimality with respect to an entire class of losses can rarely be achieved, and discuss how to find optimal solutions for specific choices of loss functions. The functionals we consider are of the form

$$\underline{T} = (T, \underline{L}),$$

where T is a one-dimensional elicitable functional with a strictly consistent loss function L, and

$$\underline{L}(P) := \inf_{x_1 \in \mathbb{R}} L(x_1, P), \tag{1}$$

where $L(x_1, P) = \int_{\mathbb{R}} L(x_1, y)\, \mathrm{d}P(y)$ is the *Bayes risk*. The example $\underline{T} = (\mathbb{E}, \mathrm{var})$ arises by choosing $L(x, y) = (x-y)^2$, and the example $\underline{T} = (q_\alpha, \mathrm{ES}_\alpha)$ is obtained by choosing $L(x, y) = (1/\alpha)\mathbb{1}\{y \le x\}(x - y) - x$, which is the piecewise linear loss known from quantile regression up to a function that only depends on y. Generally, Frongillo and Kash (2020) show that $\underline{T}$ is always 2-elicitable. Moreover, they also introduce a large class $\underline{\mathcal{L}}$ of loss functions $L(x_1, x_2, y)$ eliciting $\underline{T}$.

We show how the isotonic regression problem can be solved for $\underline{T}$. It turns out that the proposed canonical solution is generally not optimal with respect to all loss functions in $\underline{\mathcal{L}}$, but there is a fairly simple approach to check whether a given fit is simultaneously optimal. Furthermore, we show how the fit can be improved for a specific chosen loss function. In a simulation experiment, we investigate how often simultaneously optimal fits occur for the functionals $(q_\alpha, \mathrm{ES}_\alpha)$ and $(\mathbb{E}, \mathrm{var})$ and investigate the fits for a specific choice of loss function.

The article is organized as follows. Section 2 introduces necessary preliminaries on consistent loss functions including a mixture representation for loss functions in $\underline{\mathcal{L}}$. In Section 3, the isotonic regression problem for total orders is formulated and a natural solution through sequential optimization is proposed.

Then, we study the simultaneous optimality of the solution of the sequential optimization approach. Section 4 contains the numerical examples. In the Appendix, we show how our results can be generalized to partial orders.

2. Preliminaries

Following Jordan, Mühlemann and Ziegel (2020), 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 probability measure P on $\mathbb{R}$ with finite support, we define the functional T induced by an identification function V as

$$T(P) = [T^-(P), T^+(P)] \subseteq [-\infty, \infty],$$

where the lower and upper bounds are given by

$$T^-(P) = \sup\{x_1 : V(x_1, P) < 0\} \quad \text{and} \quad T^+(P) = \inf\{x_1 : V(x_1, P) > 0\},$$

using the notation $V(x_1, P) = \int_{\mathbb{R}} V(x_1, y) \,\mathrm{d}P(y)$. If a functional T is induced by an identification function, we call it *identifiable*. A broad class of functionals can be defined via their identification function, quantiles and expectiles, including the median and the mean, being just some of the most prominent examples. For other popular examples, see Jordan, Mühlemann and Ziegel (2020). The examples of quantiles and expectiles already illustrate that the functional T can take singleton-values as well as interval-values.

Recall that for a functional $\underline{T}$ taking values in $\mathbb{R}^k$, a loss function $\tilde{L} : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$ is called *consistent* if $\mathbb{E}_P \tilde{L}(t, Y) \leq \mathbb{E}_P \tilde{L}(x, Y)$ for all relevant probability distributions P, all $t \in \underline{T}(P)$ and all $x \in \mathbb{R}^k$. It is called *strictly consistent* if equality implies that $t \in \underline{T}(P)$. If $\underline{T}$ has strictly consistent loss function $\tilde{L}$, we equivalently say that $\tilde{L}$ elicits $\underline{T}$.

Theorem 1 in Frongillo and Kash (2020) states that if L is a strictly consistent loss function for T and $\underline{L}$ is the Bayes risk defined at (1), then the loss

$$\tilde{L}(x_1, x_2, y) = L'(x_1, y) + H(x_2) + h(x_2)(L(x_1, y) - x_2) \tag{2}$$

elicits $\underline{T} = (T, \underline{L})$, where $h : \mathbb{R} \to \mathbb{R}$ is any positive strictly decreasing function, $H(r) = \int_0^r h(x)\,\mathrm{d}x$, and L' is any consistent loss function for T (possibly different from L or even equal to zero). If h is merely decreasing, then $\tilde{L}$ is still a consistent loss function.

Ehm et al. (2016) pointed out that for expectiles and quantiles, any consistent loss function L' can be written as

$$L'(x_1, y) = \int_{\mathbb{R}} S_{\eta,1}(x_1, y) \,\mathrm{d}H_1(\eta), \tag{3}$$

for certain elementary (quantile or expectile) losses $S_{\eta,1}$ and a measure H_1 on $\mathbb{R}$ depending on L'; see also Lambert (2019) and the references therin. In fact, such mixtures always yield a large class $\mathcal{L}$ of consistent scoring functions for

T if it is identifiable with identification function $V(x, y)$ (Dawid, 2016; Ziegel, 2016). Then, the elementary losses are given by

$$S_{\eta,1}(x_1, y) = (\mathbb{1}\{\eta \leq x_1\} - \mathbb{1}\{\eta \leq y\})\, V(\eta, y), \tag{4}$$

where $\eta \in \mathbb{R}$. Moreover, the elementary losses are themselves consistent for T. We define

$$\mathcal{L} = \left\{ (x_1, y) \mapsto \int_{\mathbb{R}} S_{\eta,1}(x_1, y)\, \mathrm{d}H_1(\eta) : H_1 \text{ is a positive measure on } \mathbb{R} \right\}.$$

A priori, the class $\mathcal{L}$ of loss functions depends on the choice of the identification function V for T, and for simplicity we assume that we have fixed and identification function for our functional of choice throughout. However, for many functionals including all the examples mentioned above, Osband's principle for identification functions (Dimitriadis, Fissler and Ziegel, 2020, Appendix B) shows that the identification function is uniquely determined up to multiplication with a positive function that only depends on the first argument (η in (4)). This implies that the class $\mathcal{L}$ is in fact identical for any choice of identification function for T.

Note that (strict) consistency of a loss function is not altered by adding functions in y as long as they are integrable for all relevant probability measures P. Therefore, when speaking of characterizations of the class of (strictly) consistent loss functions this is always meant up to possible addition of a function in y.

If a loss function is given as a mixture of elementary losses as in (3), this may be useful when minimizing the expected loss (over some set of parameters, for example); see details for the isotonic regression problem in Section 3. Using Fubini's theorem, one can see that we can look for minimizers of the expected elementary losses and hope that these minimizers all agree, that is, there is a simultaneous minimizer for all parameters η. Then, this minimizer is automatically optimal for all scoring functions of the form (3), independently of the measure H_1. Indeed, this approach is at the heart of the characterization of all simultaneously optimal solutions to the isotonic regression problem for one-dimensional functionals in Jordan, Mühlemann and Ziegel (2020).

Using the same approach as used by Ziegel et al. (2020) to derive a mixture representation for the pair $(q_\alpha, \mathrm{ES}_\alpha)$, we derive a mixture representation for the loss functions for $\underline{T}$ of the form (2).

Lemma 2.1. *Let $L, L' \in \mathcal{L}$. Then, any consistent loss function for $\underline{T} = (T, \underline{L})$ of the form given at (2) can be written as*

$$\tilde{L}(x_1, x_2, y) = \int_{\mathbb{R}} S_{\eta,1}(x_1, y)\, \mathrm{d}H_1(\eta) + \int_{\mathbb{R}} S_{\eta,2}(x_1, x_2, y)\, \mathrm{d}H_2(\eta), \tag{5}$$

where H_1, H_2 are measures on $\mathbb{R}$, H_2 is finite on intervals of the form $(-\infty, -x_2]$, $x_2 \in \mathbb{R}$, and

$$S_{\eta,1}(x_1, y) = (\mathbb{1}\{\eta \leq x_1\} - \mathbb{1}\{\eta \leq y\})V(\eta, y)$$

$$S_{\eta,2}(x_1, x_2, y) = \mathbb{1}\{\eta \leq -x_2\}(L(x_1, y) + \eta) - \mathbb{1}\{\eta \leq 0\}\eta.$$

Conversely, any loss function of the form (5) *is consistent for* $\underline{T} = (T, \underline{L})$*. It is strictly consistent if* H_2 *puts positive mass on all open intervals.*

Proof. The consistency follows directly from Theorem 1 in Frongillo and Kash (2020). Recall that h is decreasing and nonnegative and $H(r) = \int_0^r h(x)\,\mathrm{d}x$. To see that the loss functions in (2) with loss $L' \in \mathcal{L}$ can be written as in (5), define $A := \lim_{x\to\infty} h(x) \geq 0$. Since $h \geq 0$, we can define the measure H_2 by $H_2((-\infty, t]) = h(-t) - A \geq 0$ for all $t \in \mathbb{R}$. Without loss of generality we can assume h satisfies $\lim_{x\to\infty} h(x) = 0$. Indeed, we can define $\underline{h} = h - A$ then H becomes $\underline{H}(x) = H(x) - xA$ and $\tilde{L}(x_1, x_2, y) = \tilde{\underline{L}}(x_1, x_2, y) + AL(x_1, y)$. Then, $\tilde{\underline{L}}(x_1, x_2, y) = L'(x_1, y) - AL(x_1, y) + H(x_2) + h(x_2)(L(x_1, y) - x_2)$. Thus, adding constants to h corresponds to modifying the loss function L'. Moreover, since $L, L' \in \mathcal{L}$ we have that $L' + AL \in \mathcal{L}$. Hence, we can assume that $A = 0$, then $H_2((-\infty, x]) = h(-x)$ for all $x \in \mathbb{R}$ and

$$h(x_2) = \int_{-\infty}^{-x_2} \mathrm{d}H_2(\eta).$$

Then we have

$$\int_{\mathbb{R}} S_{\eta,2}(x_1, x_2, y)\ \mathrm{d}H_2(\eta) = L(x_1, y)h(x_2) - \int_{-x_2}^{0} \eta\,\mathrm{d}H_2(\eta).$$

Integration by parts yields

$$\int_{\mathbb{R}} S_{\eta,2}(x_1, x_2, y)\ \mathrm{d}H_2(\eta) = L(x_1, y)h(x_2) - x_2 h(x_2) + H(x_2).$$

Restricting the choice of L' to $\mathcal{L}$ ensures the existence of the mixture representation for $L'(x_1, y)$. □

The following two examples discuss the mixture representations for the pairs $(q_\alpha, \mathrm{ES}_\alpha)$ and $(\mathbb{E}, \mathrm{var})$ in more detail.

Example 2.2. As mentioned in the introduction, a popular but non-elicitable risk measure is expected shortfall. In this article we adopt the sign convention used by Frongillo and Kash (2020) which is different from Fissler and Ziegel (2016); Ziegel et al. (2020).

For a given level $\alpha \in (0, 1)$, the loss function

$$L(x_1, y) = \frac{1}{\alpha}\mathbb{1}\{y \leq x_1\}(x_1 - y) - x_1$$

elicits the α-quantile $q_\alpha(P)$. The expected shortfall ES_α is the corresponding Bayes risk, that is,

$$\mathrm{ES}_\alpha(P) = \inf_{x_1 \in \mathbb{R}} L(x_1, P).$$

The elementary loss functions of Lemma 2.1 are given by

$$S_{\eta,1}(x_1,y) = (\mathbb{1}\{\eta \le x_1\} - \mathbb{1}\{\eta \le y\})(\mathbb{1}\{\eta > y\} - \alpha)$$
$$S_{\eta,2}(x_1,x_2,y) = \mathbb{1}\{\eta \le -x_2\}\left(\tfrac{1}{\alpha}\mathbb{1}\{y \le x_1\}(x_1-y) - (x_1-\eta)\right) - \mathbb{1}\{\eta \le 0\}\eta.$$

In fact, all loss functions consistent for the pair $(q_\alpha, \mathrm{ES}_\alpha)$ are of the form (2), or equivalently, (5); see Ziegel et al. (2020). Due to the different sign conventions mentioned previously, the mixture representation in Ziegel et al. (2020) corresponds to $L(x_1, -x_2, y)$ (up to normalization).

Example 2.3. The squared loss $L(x_1, y) = (x_1 - y)^2$ elicits the expectation $\mathbb{E}(P)$. The corresponding Bayes risk is the variance $\mathrm{var}(P)$. Thus, the pair $(\mathbb{E}, \mathrm{var})$ is elicitable. The elementary loss functions of Lemma 2.1 are given by

$$S_{\eta,1}(x_1,y) = (\mathbb{1}\{\eta \le x_1\} - \mathbb{1}\{\eta \le y\})(\eta - y)$$
$$S_{\eta,2}(x_1,x_2,y) = \mathbb{1}\{\eta \le -x_2\}\left((x_1-y)^2 + \eta)\right) - \mathbb{1}\{\eta \le 0\}\eta.$$

In contrast to the pair $(q_\alpha, \mathrm{ES}_\alpha)$ not all consistent loss functions for $(\mathbb{E}, \mathrm{var})$ are of this form; see Frongillo and Kash (2020, Section 3.1).

3. Isotonic regression

3.1. General results

Suppose we have pairs of observations $(z_1, y_1), \dots, (z_n, y_n)$, where $y_1, \dots, y_n$ are real-valued, the covariates $z_1, \dots, z_n$ are equipped with a total order, and $z_1 < z_2 < \cdots < z_n$. Repeated observations can easily be accommodated; see Remark 3.1 in Jordan, Mühlemann and Ziegel (2020). We aim to fit a function $\hat{g} = (\hat{g}_1, \hat{g}_2) : \{z_1, \dots, z_n\}^2 \to \mathbb{R}^2$ to these observations, such that g_1 is isotonic and models the conditional functional T given the covariates z_i, and g_2 is antitonic and models the conditional Bayes risk $\underline{L}$ given at (1) given the covariates z_i for some consistent loss function $L \in \mathcal{L}$. That is, if $z_i \le z_j$ then $\hat{g}_1(z_i) \le \hat{g}_1(z_j)$ and $\hat{g}_2(z_i) \ge \hat{g}_2(z_j)$, respectively. Considering the pair $(q_\alpha, \mathrm{ES}_\alpha)$ for example, one would be interested in an isotonic $\hat{g}_1$ and an antitonic $\hat{g}_2$ since $q_\alpha(Y_1) \le q_\alpha(Y_2)$ and $\mathrm{ES}_\alpha(Y_1) \ge \mathrm{ES}_\alpha(Y_2)$ whenever $Y_1 \le Y_2$ almost surely. Keeping this leading example in mind, we focus on the case that g_1 is isotonic, or increasing, and g_2 is decreasing, or antitonic. Adaptations of the results, where g_1 is desired to be decreasing or g_2 to be increasing are straight forward.

Following the literature on loss functions for expected shortfall, we first consider loss functions of the form (2) with $L' = 0$ (Nolde and Ziegel, 2017; Patton, Ziegel and Chen, 2019). When studying simultaneous optimality of solutions in Section 3.3, we also consider $L' \ne 0$. Let $h : \mathbb{R} \to (0, \infty)$ be decreasing with $\lim_{x\to\infty} h(x) = 0$ and $H(r) = \int_0^r h(x)\,\mathrm{d}x$. The goal is to minimize

$$\sum_{i=1}^n \tilde{L}(g_1(z_i), g_2(z_i), y_i) = \sum_{i=1}^n \Big(H(g_2(z_i)) + h(g_2(z_i))\big(L(g_1(z_i), y_i) - g_2(z_i)\big)\Big) \tag{6}$$

over all functions $g = (g_1, g_2) : \{z_1, \ldots, z_n\}^2 \to \mathbb{R}^2$ such that g_1 is increasing and g_2 is decreasing. Keeping either g_1 or g_2 fixed, we can directly give an optimal solution with respect to the other component.

Proposition 3.1. *(a) Let $g_1 : \{z_1, \ldots, z_n\} \to \mathbb{R}$ be given. Then, the optimal antitonic solution $\hat{g}_2$ of (6) with g_1 fixed is given by*

$$\hat{g}_2(z_\ell) = -\min_{j \geq \ell} \max_{i \leq j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i \leq \ell} \min_{j \geq i} -\mathbb{E}(\bar{P}_{i:j}), \quad \ell = 1, \ldots, n,$$

where $\bar{P}_{i:j}$ is the empirical distribution of $L(g_1(z_i), y_i), \ldots, L(g_1(z_j), y_j)$.

(b) Let $g_2 : \{z_1, \ldots, z_n\} \to \mathbb{R}$ be given. Then, any optimal isotonic solution $\hat{g}_1$ of (6) with g_2 fixed satisfies

$$\min_{j \geq \ell} \max_{i \leq j} T^-(P^w_{i:j}) \leq \hat{g}_1(z_\ell) \leq \max_{i \leq \ell} \min_{j \geq i} T^+(P^w_{i:j}),$$

where $P^w_{i:j}$ is the weighted empirical distribution of $y_i, \ldots, y_j$ with weights proportional to $h(g_2(z_i)), \ldots, h(g_2(z_j))$.

Proof. (a) Notice that for fixed g_1, the loss function (6) is a Bregman loss function. Moreover, $\hat{g}_2$ is antitonic if and only if $-\hat{g}_2$ is isotonic. Thus, we can solve the classical isotonic regression problem as in Jordan, Mühlemann and Ziegel (2020) for $-\hat{g}_2$ to obtain the optimal antitonic $\hat{g}_2$.

(b) Minimizing (6) for fixed g_2 is equivalent to minimizing

$$\sum_{i=1}^n h(g_2(z_i)) L(g_1(z_i), y_i).$$

Remark 3.1 and Proposition 3.6 in Jordan, Mühlemann and Ziegel (2020) yields the result. □

The bounds on the solution in Proposition 3.1 (b) are sharp in the sense that, both, the left hand side and the right hand side of the sequence of inequalities are themselves optimal solutions. However, not any increasing function $\hat{g}_1$ between those bounds is an optimal solution. A counter-example can be found in Mösching and Dümbgen (2020, Remark 2.2 and Example 2.4).

If T is singleton-valued, Proposition 3.1 yields the existence and a necessary condition for any solution to (6).

Corollary 3.2. *If T is singleton-valued, an optimal solution $\hat{g}_1, \hat{g}_2$ to (6) exists. In particular, we have*

$$\hat{g}_2(z_\ell) = -\min_{j \geq \ell} \max_{i \leq j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i \leq \ell} \min_{j \geq i} -\mathbb{E}(\bar{P}_{i:j}),$$

where $\bar{P}_{i:j}$ is the empirical distribution of $L(\hat{g}_1(z_i), y_i), \ldots, L(\hat{g}_1(z_j), y_j)$, and

$$\hat{g}_1(z_\ell) = \min_{j \geq \ell} \max_{i \leq j} T(P^w_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} T(P^w_{i:j}),$$

where $P^w_{i:j}$ is the weighted empirical distribution of $y_i, \ldots, y_j$ with weights proportional to $h(\hat{g}_2(z_i)), \ldots, h(\hat{g}_2(z_j))$.

Proof. For all solutions that are given by a min-max-representation with respect to some functional $\tilde{T}$ there exists a partition $\mathcal{Q}$ of the index set with $g(z_\ell) = \tilde{T}(Q)$, $\ell \in Q$, $Q \in \mathcal{Q}$ (Jordan, Mühlemann and Ziegel, 2020, Proposition 4.17). Since there exist only finitely many partitions of the index set $\{1, \dots, n\}$ there exist only finitely many possible solutions. Therefore, an optimal solution has to exist. In particular, $\hat{g}_1$ has to be the solution obtained from Proposition 3.1 when $\hat{g}_2$ is treated as fixed and vice versa. Otherwise we could replace $\hat{g}_1$ by the solution obtained from Proposition 3.1 to obtain a smaller loss. Similarly, we could replace $\hat{g}_2$ by the solution in Proposition 3.1 to obtain a smaller loss. □

In addition to the theoretical solutions, Proposition 3.1 suggests an algorithm for finding minimizers of (6), which roughly consists of the following steps:

1. Take g_2 constant and find the optimal $\hat{g}_1^{(1)}$.
2. Find the optimal $\hat{g}_2^{(1)}$ given $\hat{g}_1^{(1)}$.
3. Find the optimal $\hat{g}_1^{(2)}$ given $\hat{g}_2^{(1)}$.
4. Iterate steps 2 and 3 until $\hat{g}_1^{(k)} = \hat{g}_1^{(k-1)}$.

There is a problem with this algorithm if T is interval-valued, since then, the solution in part (b) of Proposition 3.1 is not unique. It turns out that it is best to choose the smallest possible solution corresponding to T^-, see Section 3.2 for details.

Fissler and Ziegel (2019) show that the expectation of consistent loss functions has no local minima. The optima in the isotonic regression case are more complex. But we believe that order sensitivity can be exploited to argue that the above algorithm can only converge to a global optimum. Numerical considerations where we perturbed the initial solutions to see whether they still converge to the same solution reinforced our suspicions that the algorithm does not converge to a saddle point. However, a rigorous mathematical proof for this conjecture is currently an open problem.

3.2. Solution to the optimization problem

In this somewhat technical section, we will show that for fixed g_2, it is best to choose

$$\hat{g}_1^-(z_\ell) := \min_{j \geq \ell} \max_{i \leq j} T^-(P^w_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} T^-(P^w_{i:j}), \tag{7}$$

where $P^w_{i:j}$ is the weighted empirical distribution of $y_{i:j}$ with weights $w_i, \dots, w_j$ proportional to $h(g_2(z_i)), \dots, h(g_2(z_j))$, to minimize (6); see Propositions 3.7 and 3.8.

The weight vector w that is described in the previous paragraph will remain fixed throughout this section. Recall that the function $\hat{g}_1^-$ is the pointwise smallest solution minimizing

$$\sum_{i=1}^n w_i L(\hat{g}_1(z_i), y_i) \tag{8}$$

over all isotonic functions $\hat{g}_1$ (Jordan, Mühlemann and Ziegel, 2020).

We denote $T^\lambda = \lambda T^- + (1-\lambda)T^+$, $\lambda \in [0,1]$, where T^- and T^+ are the lower and upper bound of T, respectively. In (7) the indices ℓ, i and j are all elements of the index set $\{1,\dots,n\}$. If we were to restrict ℓ, i and j to be elements of the subset $\{1,\dots,m\}$, $m \le n$, we would obtain an optimal solution to (8) on the subset $(z_1,y_1),\dots,(z_m,y_m)$ of the original data set, that is, with n replaced by m. In the following, we denote an optimal solution on this subset by $\hat{g}_{1;1:m}$ and by $\hat{g}_1\big|_{1:m}$ we denote the optimal solution on the original set restricted to $\{z_1,\dots,z_m\}$.

The following auxiliary result relates $\hat{g}_{1;1:m}$ to $\hat{g}_1\big|_{1:m}$ in the case where $\hat{g}_1$ is given by a min-max-representation.

Lemma 3.3. *Assume that*

$$\hat{g}_1(z_\ell) := \min_{j\ge\ell}\max_{i\le j} T^\lambda(P^w_{i:j}) = \max_{i\le\ell}\min_{j\ge i} T^\lambda(P^w_{i:j}), \quad \ell = 1,\dots,n,$$

for some $\lambda \in [0,1]$. *Then we have* $\hat{g}_1\big|_{1:m} \le \hat{g}_{1;1:m}$ *and* $\hat{g}_1\big|_{(m+1):n} \ge \hat{g}_{1;(m+1):n}$.

Proof. Notice that

$$\hat{g}_{1;1:m}(z_\ell) = \min_{\substack{j\ge\ell\\ j\le m}}\max_{i\le j} T^\lambda(P^w_{i:j}) \ge \min_{j\ge\ell}\max_{i\le j} T^\lambda(P^w_{i:j}) = \hat{g}_1(z_\ell).$$

The second statement follows with similar reasoning. □

We recall some observations made in Jordan, Mühlemann and Ziegel (2020). For fixed g_2, that is, for fixed weights w, we can minimize

$$\sum_{i=1}^n w_i \mathbb{1}\{\eta \le \hat{g}_1(z_i)\} V(\eta, y_i), \quad \text{for all } \eta \in \mathbb{R}, \tag{9}$$

to obtain an optimal solution to (6), or equivalently, to (8). The crucial insight of Jordan, Mühlemann and Ziegel (2020) is that an optimal isotonic solution to (8) is not only optimal for L but also for any other $L' \in \mathcal{L}$. Therefore, we can equivalently find a solution that is optimal for all elementary losses given in Lemma 2.1, which leads to (9). Because we want $\hat{g}_1$ to be isotonic, this means that for a given $\eta \in \mathbb{R}$ we have to find an index $\ell \in \{1\dots,n+1\}$ that minimizes

$$\sum_{i=\ell}^n w_i V(\eta, y_i). \tag{10}$$

For $\ell = n+1$, recall that an empty sum is zero. The search for the optimal index ℓ needs to be conducted for every $\eta \in \mathbb{R}$. For $\eta \in \mathbb{R}$, we denote the set of indices minimizing (10) by $I_{1:n}(\eta)$.

Lemma 3.4. *We have that* $I_{1:n}(\eta) \cap \{1,\dots,m+1\} \subseteq I_{1:m}(\eta)$, *where* $I_{1:m}(\eta)$ *is the set of minimizing indices for the isotonic regression problem* (8) *on the subsample* $(z_1,y_1),\dots,(z_m,y_m)$.

Proof. Let $\ell \in I_{1:n}(\eta) \cap \{1, \dots, m\}$ for some $\eta \in \mathbb{R}$. Therefore, the function

$$t_\eta : \{1, \dots, n+1\} \to \mathbb{R}, x \mapsto \sum_{i=x}^{n} w_i V(\eta, y_i)$$

has a minimum at ℓ. We can write

$$\sum_{i=\ell}^{n} w_i V(\eta, y_i) = \sum_{i=\ell}^{m} w_i V(\eta, y_i) + \sum_{i=m+1}^{n} w_i V(\eta, y_i).$$

Hence, $t_\eta\big|_{1:m}$ has also a minimum at ℓ and thus $\ell \in I_{1:m}(\eta)$. If t_η has a minimum at $\ell = m+1$ then

$$t_\eta(x) - \sum_{i=m+1}^{n} w_i V(\eta, y_i) \geq 0$$

with equality for $x = m+1$. Thus, $I_{1:n}(\eta) \cap \{1, \dots, m+1\} \subseteq I_{1:m}(\eta)$. □

The next result shows that if $\hat{g}_1$ is an optimal solution to the isotonic regression problem (8) with $\hat{g}_1(z_m) < \hat{g}_1(z_{m+1})$ then $\hat{g}_1\big|_{1:m}$ is an optimal solution to the isotonic regression problem (8) on the subsample $(z_1, y_1), \dots, (z_m, y_m)$.

Corollary 3.5. *Let $\hat{g}_1$ be an optimal solution to (8) with $\hat{g}_1(z_m) < \hat{g}_1(z_{m+1})$. Then we have that $\hat{g}_1\big|_{1:m}$ is an optimal solution to (8) on the subsample (z_1, y_1), $\dots$, (z_m, y_m).*

Proof. It follows from Jordan, Mühlemann and Ziegel (2020, Proposition 3.5) that the optimal solution $\hat{g}_1$ is in one-to-one correspondence with increasing, left-continuous functions $\iota : \mathbb{R} \to \{1, \dots, n+1\}$ with $\iota(\eta) \in I_{1:n}(\eta)$, for all $\eta \in \mathbb{R}$, in the sense that

$$\inf\{\eta : \iota(\eta) > \ell\} = \hat{g}_1(z_\ell) = \max\{\eta : \iota(\eta) \leq \ell\}.$$

The restricted function $\iota\big|_{1:m} : \mathbb{R} \to \{1, \dots, m+1\}$ is left-continuous, increasing and satisfies

$$\inf\{\eta : \iota\big|_{1:m}(\eta) > \ell\} = \hat{g}_1\big|_{1:m}(z_\ell) = \max\{\eta : \iota\big|_{1:m}(\eta) \leq \ell\}.$$

Moreover, $\iota\big|_{1:m}(\eta) \in I_{1:n}(\eta) \cap \{1, \dots, m+1\}$, for all $\eta \in \mathbb{R}$. Lemma 3.4 implies that $\iota\big|_{1:m}(\eta) \in I_{1:m}(\eta)$, for all $\eta \in \mathbb{R}$. Thus, $\hat{g}_1\big|_{1:m}$ is an optimal solution to (8) on the subsample $(z_1, y_1), \dots, (z_m, y_m)$. □

We now would like to show that for fixed weights the solution

$$\hat{g}_1^-(z_\ell) = \min_{j \geq \ell} \max_{i \leq j} T^-(P^w_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} T^-(P^w_{i:j})$$

is most likely to minimize (6) out of all possible solutions to (8). An intuition behind this statement is obtained by combining Lemma 3.3 with Proposition

3.6 from Jordan, Mühlemann and Ziegel (2020). Lemma 3.3 yields $\hat{g}_1^-\big|_{(m+1):n} \geq \hat{g}^-_{1,(m+1):n}$. Proposition 3.6 of Jordan, Mühlemann and Ziegel (2020), on the other hand, implies that any optimal solution $\hat{g}_{1,(m+1):n}$ on $(z_{m+1}, y_{m+1}), \dots, (z_n, y_n)$ has to satisfy $\hat{g}^-_{1,(m+1):n} \leq \hat{g}_{1,(m+1):n} \leq \hat{g}^+_{1,(m+1):n}$, where $\hat{g}^+_{1,(m+1):n}$ is the upper bound given by Proposition 3.1 (b) when considering the subsample $\{z_{m+1}, \dots, z_n\}$. Thus, $\hat{g}_1^-$ has the highest chance to lie between those bounds.

To prove this formally the order sensitivity of loss functions is needed. We recall the definition given in Steinwart et al. (2014).

Definition 3.6. Let $\mathcal{P}$ be a class of probability distributions. A loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is said to be *$\mathcal{P}$-order sensitive for T*, if the image of T is an interval, and for all $P \in \mathcal{P}$ and all $t_1, t_2 \in \mathbb{R}$ with either $t_2 < t_1 \leq T^-(P)$ or $T^+(P) \leq t_1 < t_2$, we have $L(t_1, P) < L(t_2, P)$.

It follows directly from the definition that order sensitive loss functions are consistent. The reverse holds under weak regularity conditions on the functional; see Lambert (2019, Proposition 11). The loss functions in class $\mathcal{L}$ are order-sensitive because they are defined via an oriented identification function and a positive measure H_1 (Steinwart et al., 2014, Theorem 7). Thus, the loss function L in the following proposition is order sensitive. This observation is crucial for the final step in the proof.

Proposition 3.7. *Let $\hat{g}_1^-$ be given by (7), and let $\hat{g}_1$ be any other solution to (8). Then, for all antitonic g_2', we have*

$$\sum_{i=1}^{n} \tilde{L}(\hat{g}_1^-(z_i), g_2'(z_i), y_i) \leq \sum_{i=1}^{n} \tilde{L}(\hat{g}_1(z_i), g_2'(z_i), y_i).$$

Proof. For each $\hat{g}_1$, we have a partition $\mathcal{Q}$ of the index set such that

$$\hat{g}_1(z_i) = \hat{g}_1(z_j) \quad \text{for all } i, j \in Q,\ Q \in \mathcal{Q}.$$

We denote by Q_m the partition element corresponding to $\hat{g}_1$ containing m, and by Q_m^- we denote the partition element corresponding to $\hat{g}_1^-$ containing m. Recall that we assumed that $L' = 0$. Therefore, by Lemma 2.1, it suffices to show that for all $\eta \in \mathbb{R}$, and for all antitonic g_2'

$$\sum_{i=1}^{n} S_{\eta,2}(\hat{g}_1^-(z_i), g_2'(z_i), y_i) \leq \sum_{i=1}^{n} S_{\eta,2}(\hat{g}_1(z_i), g_2'(z_i), y_i),$$

since $S_{1,\eta}$ only occurs in the mixture representation of L'. For the latter, it suffices to show that for all $m \leq n$

$$\sum_{\ell=m}^{n} L(\hat{g}_1^-(z_\ell), y_\ell) \leq \sum_{\ell=m}^{n} L(\hat{g}_1(z_\ell), y_\ell). \tag{11}$$

Since $\hat{g}_1$ is another solution to (8), $\hat{g}_1$ can only jump in ℓ with $\ell \in \cup_\eta I_{1:n}(\eta)$. In particular, we have

$$\sum_{\ell=1}^{n} L(\hat{g}_1^-(z_\ell), y_\ell) = \sum_{\ell=1}^{n} L(\hat{g}_1(z_\ell), y_\ell).$$

In the following, we will prove the converse to (11), that is, for all $m \leq n$ we have

$$\sum_{\ell=1}^{m} L(\hat{g}_1(z_\ell), y_\ell) \leq \sum_{\ell=1}^{m} L(\hat{g}_1^-(z_\ell), y_\ell). \tag{12}$$

If $m = \max Q_m$, it follows from Corollary 3.5 that $\hat{g}_1\big|_{1:m}$ is optimal on (z_1, y_1), $\dots$, (z_m, y_m) and therefore (12) holds.

For $m \neq \max Q_m$, we distinguish two cases.

Case 1: If $m = \max Q_m^-$, it follows from Lemma 3.3 and Proposition 3.1 that

$$\hat{g}_1^-\big|_{1:m} = \hat{g}_{1;1:m}^- \leq \hat{g}_1\big|_{1:m} \leq \hat{g}_1^+\big|_{1:m} \leq \hat{g}_{1;1:m}^+.$$

By Lemma 3.4 we have $I_{1:n}(\eta) \cap \{1, \dots, m+1\} \subseteq I_{1:m}(\eta)$ for all $\eta \in \mathbb{R}$. Hence, $\iota\big|_{1:m}(\eta) \in I_{1:m}(\eta)$ for all $\eta \in \mathbb{R}$, where $\iota : \mathbb{R} \to \{1, \dots, n+1\}$ is the function imposing the score minimizing-indices corresponding to $\hat{g}_1$. Thus, Proposition 3.5 in Jordan, Mühlemann and Ziegel (2020) implies that $\hat{g}_1\big|_{1:m}$ is an optimal solution to the isotonic regression problem on $(z_1, y_1), \dots, (z_m, y_m)$.

Case 2: Consider the case $m \neq \max Q_m^-$ and let $j = \max(\min Q_m, \min Q_m^-)$. It follows from the previous considerations that $\hat{g}_1$ is optimal up to $j-1$ in the sense that it is a minimizer on $(z_1, y_1), \dots, (z_{j-1}, y_{j-1})$. We know that $\hat{g}_1^-\big|_{1:m} \leq \hat{g}_{1;1:m}^-$ so if $\hat{g}_1\big|_{1:m} \geq \hat{g}_{1;1:m}^-$ we can conclude with the same reasoning as in case 1.

Otherwise, let $j_0 \geq j$ be the minimal index with $\hat{g}_{1;1:m}^-(z_{j_0}) > \hat{g}_1\big|_{1:m}(z_{j_0})$. Clearly $j_0 \in Q_m$ and hence $\hat{g}_1$ is constant on $\{j, \dots, j_0\}$. Moreover, if $j_0 > j$ then for all $\ell \in \{1, \dots, j_0 - 1\}$ we have that

$$\hat{g}_{1;1:m}^-(z_\ell) = \hat{g}_{1;1:(j_0-1)}^-(z_\ell) \leq \hat{g}_1(z_\ell) \leq \hat{g}_1^+\big|_{1:(j_0-1)}(z_\ell) \leq \hat{g}_{1;1:(j_0-1)}^+(z_\ell),$$

implying that $\hat{g}_1$ is in fact optimal up to $j_0 - 1$. Of course, if $j_0 = j$, we already know that $\hat{g}_1$ is optimal up to $j_0 - 1$, since we know that $\hat{g}_1$ is optimal up to $j - 1$ from our previous considerations. Thus, it remains to check what happens for $\ell \in \{j_0, \dots, m\}$.

For $\ell \in \{j_0, \dots, m\}$ we have $\hat{g}_1^-(z_\ell) = c^- \leq c = \hat{g}_1(z_\ell) < \hat{g}_{1;1:m}^-(z_\ell)$ for some constants c^- and c.

Denote by $Q_{s;1:m}^-, \dots, Q_{r;1:m}^-$ the partition elements of $\hat{g}_{1;1:m}^-$ on $\{j_0, \dots, m\}$. Then, for $k \in \{s, \dots, r\}$ we have

$$\sum_{\ell \in Q_{k;1:m}^-} L(\hat{g}_{1;1:m}^-, y_\ell) \leq \sum_{\ell \in Q_{k;1:m}^-} L(c, y_\ell) \leq \sum_{\ell \in Q_{k;1:m}^-} L(c^-, y_\ell)$$

since $\hat{g}_{1;1:m}^-$ is constant each $Q_{k;1:m}^-$ and L is order-sensitive. Therefore, (12) is fulfilled. $\square$

Finally, we have all necessary results to see that $\hat{g}_1^-$ is indeed our best bet. Define

$$\hat{g}_1^-(z_\ell; w) := \min_{j\geq\ell}\max_{i\leq j} T^-(P^w_{i:j}) = \max_{i\leq\ell}\min_{j\geq i} T^-(P^w_{i:j})$$
$$\hat{g}_2^-(z_\ell; \hat{g}_1^-) := -\min_{j\geq\ell}\max_{i\leq j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i\leq\ell}\min_{j\geq i} -\mathbb{E}(\bar{P}_{i:j}),$$

where $\bar{P}_{i:j}$ is the empirical distribution of $L(\hat{g}_1^-(z_i), y_i), \ldots, L(\hat{g}_1^-(z_j), y_j)$ and $P^w_{i:j}$ is the weighted empirical distribution of $y_i, \ldots, y_j$ with weights w.

Proposition 3.8. *Assume that there exist $\hat{g}_1, \hat{g}_2 \colon \{z_1, \ldots, z_n\} \to \mathbb{R}$ minimizing (6), then $\hat{g}_1^-(\cdot; h(\hat{g}_2)), \hat{g}_2^-(\cdot; \hat{g}_1^-(\cdot; h(\hat{g}_2)))$ are also minimizers.*

Proof. Clearly the pair $\hat{g}_1(\cdot), \hat{g}_2(\cdot)$ has to satisfy the restrictions imposed by Proposition 3.1 as otherwise they would not be optimal. Proposition 3.7 implies that the pair $\hat{g}_1^-(\cdot; h(\hat{g}_2)), \hat{g}_2(\cdot)$ is also a minimizing pair to (6). Finally applying part (a) of Proposition 3.1 we can conclude that $\hat{g}_2(\cdot) = \hat{g}_2^-(\cdot; \hat{g}_1^-(\cdot; h(\hat{g}_2)))$. □

3.3. Simultaneously optimal solutions

A simultaneously optimal solution $\hat{g}_1, \hat{g}_2$ has to minimize the expected elementary losses

$$\frac{1}{n}\sum_{i=1}^{n} S_{\eta,1}(g_1(z_i), y_i) \tag{13}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} S_{\eta,2}(g_1(z_i), g_2(z_i), y_i) \tag{14}$$

for all $\eta \in \mathbb{R}$ among all increasing functions $g_1 : \{z_1, \ldots, z_n\} \to \mathbb{R}$ and all decreasing functions $g_2 : \{z_1, \ldots, z_n\} \to \mathbb{R}$. The expected elementary score (13) is minimized for all $\eta \in \mathbb{R}$ if and only if $\hat{g}_1$ is an optimal isotonic solution with respect to T characterized in Jordan, Mühlemann and Ziegel (2020). Thus, there can only exist a simultaneously optimal solution if for one such $\hat{g}_1$ there exists $\hat{g}_2 : \{z_1, \ldots, z_n\} \to \mathbb{R}$ decreasing so that the pair $\hat{g}_1, \hat{g}_2$ minimizes (14) for all $\eta \in \mathbb{R}$.

The proof of Proposition 3.7 suggests that for any $m \leq n$

$$\sum_{i=m}^{n} L(\hat{g}_1^-(z_i), y_i) \leq \sum_{i=m}^{n} L(\hat{g}_1(z_i), y_i)$$

with equality whenever $m = n$. Note that minimizing (14) for all $\eta \in \mathbb{R}$ is equivalent to minimizing

$$\sum_{i=1}^{n} \mathbb{1}\{\eta \leq -g_2(z_i)\} L(g_1(z_i), y).$$

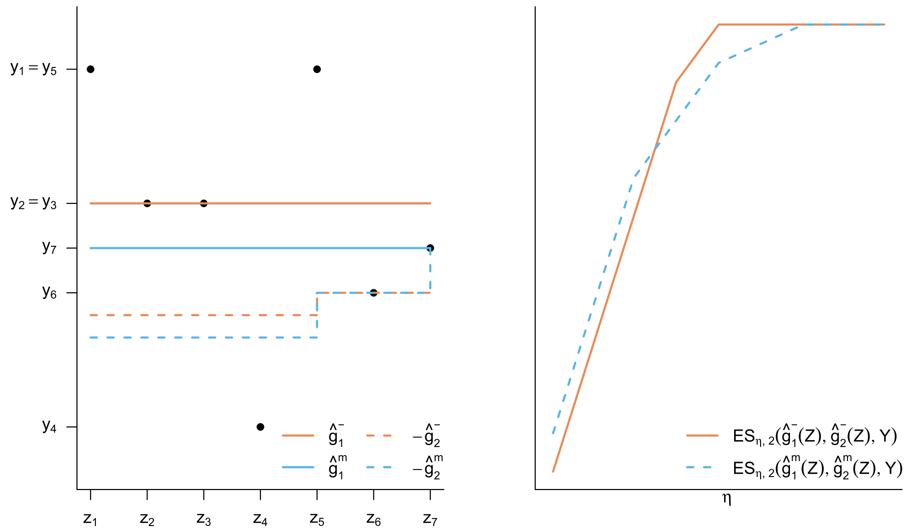

FIG 1. *Specific sample of seven data points (black) on the left, such that for* $\underline{T} = (q_{0.5}, \mathrm{ES}_{0.5})$, $\hat{g}_1^-|_{5:7}$ *is not an optimal isotonic solution on* (z_5, y_5), (z_6, y_6), (z_7, y_7) *but* $\hat{g}_2^-(z_4) > \hat{g}_2^-(z_5)$. *The function* $\hat{g}_1^m$ *is not an optimal isotonic solution to the global optimization problem. The Murphy diagram (Ehm et al., 2016) (plot of expected elementary scores) on the right shows that there are values of* η *where* $\hat{g}_1^m$, $\hat{g}_2^m$ *has smaller expected loss than* $\hat{g}_1^-$, $\hat{g}_2^-$.

Thus, a pair $\hat{g}_1, \hat{g}_2$ can only be simultaneously optimal if $\hat{g}_1\big|_{m:n}$ is an optimal isotonic solution on $(z_m, y_m), \dots, (z_n, y_n)$ for all $m \in \{1, \dots, n\}$ with $\hat{g}_2(z_{m-1}) > \hat{g}_2(z_m)$. If this is not the case for some $m \in \{1, \dots, n\}$, we can find $\hat{g}_1^m$ such that the pair $\hat{g}_1^m, \hat{g}_2^m$, where $\hat{g}_2^m$ is the corresponding solution obtained via Proposition 3.1, dominates $\hat{g}_1, \hat{g}_2$ for all $\eta \in \mathbb{R}$ with $\hat{g}_2(z_{m-1}) < \eta \leq \hat{g}_2(z_{m-1})$. But inevitably this solution performs worse for other $\eta \in \mathbb{R}$, especially for $\eta \leq -g_2(z_1)$. Figure 1 displays a data example where a simultaneously optimal solution does not exist because there exists some index m with $\hat{g}_2(z_{m-1}) > \hat{g}_2(z_m)$ but $\hat{g}_1^-\big|_{m:n}$ is not an optimal isotonic solution on $(z_m, y_m), \dots, (z_n, y_n)$. The previous considerations are summarized by the following proposition.

Proposition 3.9. *A simultaneously optimal solution exists if and only if* $\hat{g}_1^-\big|_{m:n}$ *is an optimal solution on* $(z_m, y_m), \dots, (z_n, y_n)$ *for all* $m \in \{2, \dots, n\}$ *such that* $\hat{g}_2^-(z_{m-1}) > \hat{g}_2^-(z_m)$ *and* $\hat{g}_1^-(z_{m-1}) = \hat{g}_1^-(z_m)$.

The condition $\hat{g}_1^-(z_{m-1}) = \hat{g}_1^-(z_m)$ arises because the only critical indices are those where $\hat{g}_2^-$ has a jump but $\hat{g}_1^-$ has not. In the case where $\hat{g}_1^-$ has a jump too, $\hat{g}_1^-$ is already an optimal solution on $\{m, \dots, n\}$.

Proposition 3.9 supplies us with a criterion to check for simultaneous optimality. The approach is to first calculate

$$\hat{g}_1^-(z_\ell) := \min_{j \geq \ell} \max_{i \leq j} T^-(P_{i:j}) = \max_{i \leq \ell} \min_{j \geq i} T^-(P_{i:j}),$$
$$\hat{g}_2^-(z_\ell) := -\min_{j \geq \ell} \max_{i \leq j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i \leq \ell} \min_{j \geq i} -\mathbb{E}(\bar{P}_{i:j}),$$

with $\bar{P}$ as defined in Proposition 3.1. In a second step, for each $m \geq 2$ with $\hat{g}_2^-(z_{m-1}) > \hat{g}_2^-(z_m)$ and $\hat{g}_1^-(z_{m-1}) = \hat{g}_1^-(z_m)$ one has to check whether $\hat{g}_1^-\big|_{m:n}$ is an optimal solution on the subset $(z_m, y_m), \dots, (z_n, y_n)$. To check whether $\hat{g}_1^-\big|_{m:n}$ remains optimal we can compare the expected elementary score for $\hat{g}_1^-\big|_{m:n}$ to the one of $\hat{g}_{1;m:n}^-$. If $\hat{g}_1^-\big|_{m:n}$ remains optimal for each $m \geq 2$ with $\hat{g}_2^-(z_{m-1}) > \hat{g}_2^-(z_m)$ and $\hat{g}_1^-(z_{m-1}) = \hat{g}_1^-(z_m)$, then the solution $(\hat{g}_1^-, \hat{g}_2^-)$ is indeed simultaneously optimal.

For bivariate functionals T with two elicitable components there always exists a subclass $\mathcal{L}_2$ of consistent loss functions $L(x_1, x_2, y)$ that are separable in the sense that $L(x_1, x_2, y) = L_1(x_1, y) + L_2(x_1, y)$. Solving the isotonic regression problem simultaneously over all $L \in \mathcal{L}_2$ can be split into two independent optimization problems. In this case Jordan, Mühlemann and Ziegel (2020) provide all necessary tools for a complete characterization of all solutions. But not all consistent loss functions lie necessarily in $\mathcal{L}_2$. If T is a vector of moments this can be seen in Proposition 4.11 in Fissler and Ziegel (2019). In the case where T is a vector of quantiles, however, $\mathcal{L}_2$ comprises all consistent losses (Fissler and Ziegel, 2016, Proposition 4.2) explaining some of the optimality properties of the IDR introduced by Henzi, Ziegel and Gneiting (2019). Thus, when considering functionals with elicitable components one can reach simultaneous optimality at least with respect to the class $\mathcal{L}_2$. When considering functionals with elicitation complexity greater than one however, there are no separable consistent loss functions, so that possibly no simultaneous optimum exists.

4. Numerical experiments

We let

$$\begin{aligned}\hat{g}_1^-(z_\ell) &:= \min_{j\geq\ell}\max_{i\leq j} T^-(P_{i:j}) = \max_{i\leq\ell}\min_{j\geq i} T^-(P_{i:j}),\\ \hat{g}_2^-(z_\ell) &:= -\min_{j\geq\ell}\max_{i\leq j} -\mathbb{E}(\bar{P}_{i:j}) = -\max_{i\leq\ell}\min_{j\geq i} -\mathbb{E}(\bar{P}_{i:j}).\end{aligned}$$

In this section we investigate how often simultaneous optimality occurs and the number of iterations needed to obtain an optimal solution for a specific loss function, whenever the solution $\hat{g}_1^-, \hat{g}_2^-$ is not simultaneously optimal. We consider the two prominent examples $(q_\alpha, \mathrm{ES}_\alpha)$ and $(\mathbb{E}, \mathrm{var})$ in the simulations.

First, let us examine what we would expect to result from those simulations in terms of simultaneous optimality. In Section 3.2, we saw that simultaneous optimality is attained whenever $\hat{g}_1^-\big|_{m:n}$ remains an optimal solution for all $m \in \{2, \dots, n\}$ with $\hat{g}_2^-(z_{m-1}) > \hat{g}_2^-(z_m)$. Clearly, this requirement is fulfilled as long as $\hat{g}_2^-$ jumps at the same point as $\hat{g}_1^-$. Naturally, the more jumps $\hat{g}_1^-$ has, or equivalently the less pooling was required, the higher are the chances for simultaneous optimality, in that there are not many additional restrictions left to be imposed by $\hat{g}_2^-$. Thus, the less the isotonicity constraint is violated in the data the higher the chances for the pair $(\hat{g}_1^-, \hat{g}_2^-)$ to be simultaneously optimal. Only considering the impact of $\hat{g}_1^-$, we would expect the chance for simultaneous

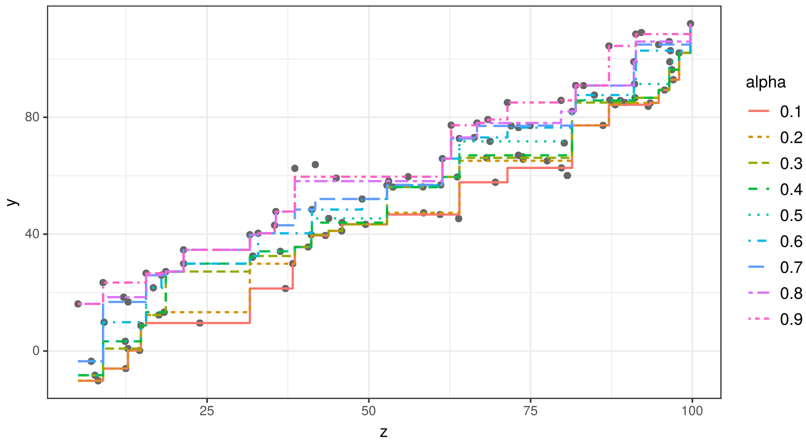

FIG 2. *For a set of* $n = 100$ *data points and the pair* $(q_\alpha, \mathrm{ES}_\alpha)$ *the optimal fit* $\hat{g}_1^-$ *was drawn for each* $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

optimality to decrease with increasing variance in the data. Moreover, for fixed variance we would expect the chance of simultaneous optimality to decrease with increasing sample size, because the chance for necessary pooling increases.

Concerning the impact of $\hat{g}_2^-$, we have seen in Proposition 3.1 that $\hat{g}_2^-$ is fitted to the transformed data points $(z_1, L(\hat{g}_1^-(z_1), y_1)), \ldots, (z_n, L(\hat{g}_1^-(z_n), y_n))$, where the transformed y-values depend on the loss L of y_ℓ and $\hat{g}_1^-(z_\ell)$. The order sensitivity of the loss function ensures that the transformation $L(\hat{g}_1^-(z_\ell), y_\ell)$ takes larger values when $\hat{g}_1^-(z_\ell)$ and y_ℓ are far apart and smaller values when they are close. Thus, if small modifications are necessary to obtain $\hat{g}_1$, then we would expect the transformed data to be approximately constant. The outcome of the transformation however depends on how the loss L weighs the differences.

The setup for the simulations was the following: For the pair $(q_\alpha, \mathrm{ES}_\alpha)$ we aimed to optimally fit an increasing function $\hat{g}_1^-$ and decreasing function $\hat{g}_2^-$ to simulated data sets. We drew n points z_ℓ independently and uniformly from $[0, 100]$. The corresponding y-value was $y_\ell = z_\ell + \epsilon_\ell$ where $\epsilon_\ell \sim \mathcal{N}(0, \sigma^2)$ are independent of each other and independent of z_ℓ. We let $n \in \{10, 100, 500, 1000\}$ and $\sigma \in \{3, 10, 20, 30\}$ and we repeated the experiment $M = 1000$ times to count the number of times simultaneous optimality occurred. To investigate whether the results differ depending on the level α, we calculated $\hat{g}_1^-$ and $\hat{g}_2^-$ for each data set for all $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. For a specific data set, Figure 2 shows the fits $\hat{g}_1^-$ for all levels α, and Figure 3 contains the corresponding fits $-\hat{g}_2^-$.

For the pair $(\mathbb{E}, \mathrm{var})$ we aimed to optimally fit two increasing functions $\hat{g}_1^-$ and $\hat{g}_2^-$ to simulated data sets. All our results can clearly be adapted to this case. Thus, again we drew n points z independently and uniformly from $[0, 100]$. The corresponding y-value was $y_\ell = z_\ell + \epsilon_\ell$ where $\epsilon_\ell \sim \mathcal{N}(0, c\ell/\sqrt{n})$ were independent. We let $n \in \{10, 20, 40, 60, 80, 100, 500, 1000\}$ and $c \in \{0.5, 1, 3, 6\}$

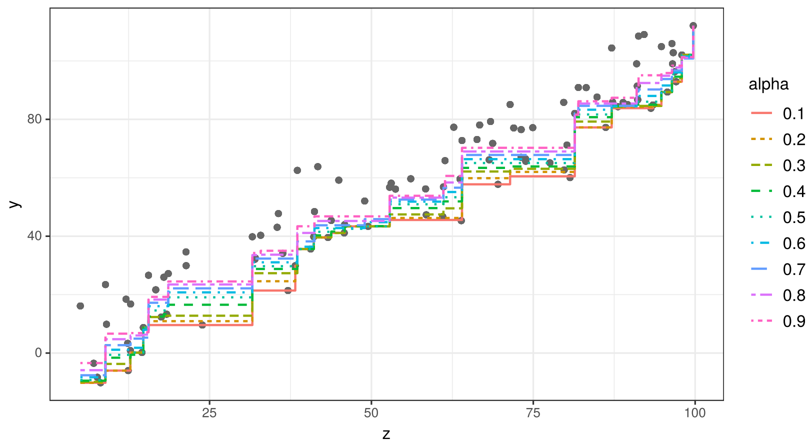

FIG 3. *For the same choice of $n = 100$ data points as in Figure 2 the corresponding fits $\hat{g}_2^-$ are calculated and $-\hat{g}_2^-$ is displayed for each $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.*

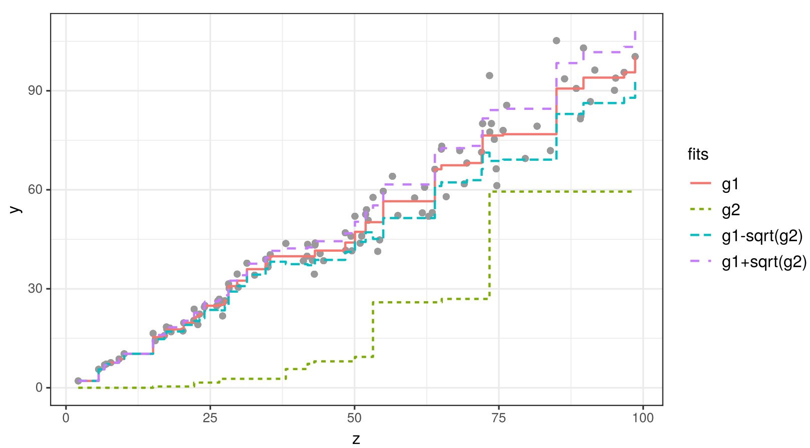

FIG 4. *For a sample of $n = 100$ data points and the pair $(\mathbb{E}, \mathrm{var})$ the optimal fit $\hat{g}_1$ is drawn in red and $\hat{g}_2$ is in green. Moreover, $\hat{g}_1 - \sqrt{\hat{g}_2}$ and $\hat{g}_1 + \sqrt{\hat{g}_2}$ are drawn in blue and pink, respectively*

and then generated $M = 1000$ data sets and calculated the corresponding fits $\hat{g}_1^-$ and $\hat{g}_2^-$. Figure 4 contains the fits $\hat{g}_1^-$ and $\hat{g}_2^-$ for a specific data set.

Using the criterion in Proposition 3.9, we counted how many times simultaneous optimality occurred. Table 1 contains the results for the pair $(q_\alpha, \mathrm{ES}_\alpha)$. The percentage of times simultaneous optimality is reached is displayed. The results confirm our expectations. With increasing sample size and increasing variance the percentage decreases drastically. The reason that not all levels α are equally affected is due to the different weights that L imposes depending on

TABLE 1. *Percentage of times simultaneous optimality occurred for* $(q_\alpha, \mathrm{ES}_\alpha)$ *for each combination of sample size* n*, standard deviation* σ*, and level* α*.*

		$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
$n = 10$	$\sigma = 3$	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.94	0.94
	$\sigma = 10$	1.00	1.00	1.00	1.00	0.98	0.96	0.79	0.70	0.69
	$\sigma = 20$	1.00	1.00	1.00	0.98	0.94	0.92	0.69	0.58	0.50
	$\sigma = 30$	1.00	1.00	0.98	0.97	0.88	0.88	0.64	0.53	0.44
$n = 100$	$\sigma = 3$	1.00	0.97	0.60	0.48	0.14	0.13	0.00	0.00	0.00
	$\sigma = 10$	0.96	0.53	0.16	0.08	0.01	0.01	0.00	0.00	0.00
	$\sigma = 20$	0.80	0.31	0.11	0.06	0.01	0.02	0.00	0.00	0.00
	$\sigma = 30$	0.70	0.27	0.12	0.06	0.02	0.02	0.00	0.00	0.00
$n = 500$	$\sigma = 3$	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 10$	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 20$	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 30$	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$n = 1000$	$\sigma = 3$	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 20$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\sigma = 30$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2
Percentage of times simultaneous optimality occurred for $(\mathbb{E}, \mathrm{var})$ *for each combination of sample size* n *and constant* c*.*

	$c = 0.5$	$c = 1$	$c = 3$	$c = 6$
$n = 10$	0.98	0.95	0.79	0.57
$n = 20$	0.88	0.73	0.33	0.14
$n = 40$	0.53	0.22	0.02	0.00
$n = 60$	0.21	0.06	0.00	0.00
$n = 80$	0.07	0.02	0.00	0.00
$n = 100$	0.02	0.00	0.00	0.00
$n = 500$	0.00	0.00	0.00	0.00
$n = 1000$	0.00	0.00	0.00	0.00

the level α.

The results for the pair $(\mathbb{E}, \mathrm{var})$ in Table 2 also confirm our expectations. The reason why the percentage in this case decreases even more rapidly is that the expectation $\mathbb{E}$ is less robust when it comes to removing data from a partition element than the quantile q_α is.

Simultaneous optimality is usually not attainable. In these cases, we have to choose a specific loss function to solve the isotonic regression problem. It is natural to ask, how different these solutions are compared to our candidate for simultaneous optimality.

For both examples, we choose two different functions h and count the number of iterations the algorithm needed to get from the candidate for simultaneous optimality to a potential optimal solution for the specific loss. For the pair $(q_\alpha, \mathrm{ES}_\alpha)$, we considered the $(1/2)$-homogeneous loss from Nolde and Ziegel (2017). It arises when choosing h in (2) as $h_1(x) = 1/(2\sqrt{x})$. We also considered $h_2(x) = \exp(-x)$. The iteration was stopped when the loss given by (6) did not improve by more that 10^{-10}. For both loss functions, almost no adjustments were necessary with a maximum average number of iterations for h_1 of 0.11 when $\sigma = 30$ and $\alpha = 0.3$, and for h_2 of 0.05 when $\sigma = 30$ and $\alpha = 0.1$. For most combinations of σ and α, the average number of iterations was zero for both loss functions which is why detailed results are not displayed. This suggests that although the candidate for simultaneous optimality is not simultaneously optimal, it still is optimal with respect to some losses.

For the pair $(\mathbb{E}, \mathrm{var})$, we chose functions $h_1(x) = 1/(x + 0.1)$ and $h_2(x) = \exp(-x/50 + 0.1)$. The reason for dividing by 50 was the scale of the weights to avoid numerical issues. The summand $+0.1$ was to avoid weights of zero. Again, the iteration was stopped when the loss given by (6) did not improve by more that 10^{-10}. Figure 5 displays the corresponding solutions obtained for a specific data set. The average number of iterations is displayed in Table 3. Here, the situation is different. Given a specific loss function for the pair $(\mathbb{E}, \mathrm{var})$, the global loss may decrease through adaptations of the optimal solution for $\mathbb{E}$ alone.

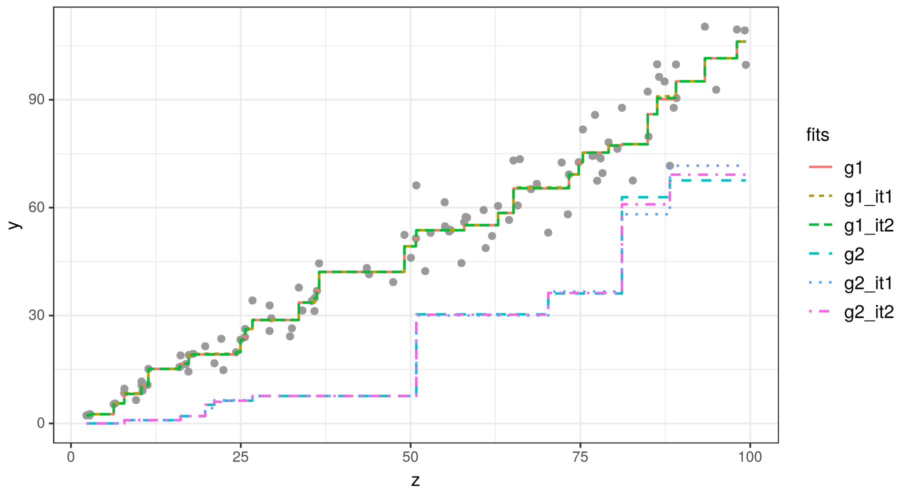

Fig 5. *For a specific sample of size* 100 *the original fits (g1 and g2) are displayed in red and light blue respectively. The light green and dark blue fits (g1_it1 and g2_it1) correspond to the iterated versions of g_1 and g_2, respectively, with respect to the weight function h_1. Finally, the dark green and the pink fits (g1_it2 and g2_it2) correspond to the iterated versions of g_1 and g_2, respectively, with respect to h_2. For h_1 the number of iterations was* 13 *and for h_2 a total of* 5 *iterations were necessary.*

Table 3
The average number of iterations are displayed for the two weight functions h_1, h_2 considered for the pair $(\mathbb{E}, \mathrm{var})$.

		$c = 0.5$	$c = 1$	$c = 3$	$c = 6$
$n = 10$	h_1	0.07	0.24	1.15	2.77
	h_2	0.01	0.06	0.94	2.79
$n = 20$	h_1	0.64	1.95	5.85	9.41
	h_2	0.09	0.66	4.05	7.78
$n = 40$	h_1	3.53	7.43	12.51	13.69
	h_2	0.59	2.42	8.70	10.81
$n = 60$	h_1	7.19	10.04	13.05	14.30
	h_2	1.53	4.27	10.76	11.88
$n = 80$	h_1	9.73	11.32	14.27	15.05
	h_2	2.11	5.25	11.44	12.40
$n = 100$	h_1	10.52	12.48	13.95	14.45
	h_2	2.90	6.52	13.12	12.73
$n = 500$	h_1	10.54	11.76	13.77	13.58
	h_2	6.68	11.05	14.40	8.78
$n = 1000$	h_1	9.16	10.05	11.37	12.58
	h_2	7.91	12.04	12.19	3.32

Appendix A: Generalizations to partial orders

The results in this article can be generalized to partially ordered covariate sets. Let distribution P be the distribution of the random vector $(Z, Y) \in \mathcal{Z} \times \mathbb{R}$, where $\mathcal{Z}$ is a finite partially ordered set. We denote the partial order by $\preceq$. The distribution P can be the empirical distribution of a sample with values but it does not have to be. However, we assume that $P(\{z\} \times \mathbb{R}) > 0$ for all $z \in \mathcal{Z}$. We aim now to minimize the criterion

$$\int_{\mathcal{Z}\times\mathbb{R}} \tilde{L}(g_1(z), g_2(z), y)\, P(\,\mathrm{d}z,\, \mathrm{d}y)$$
$$= \int_{\mathcal{Z}\times\mathbb{R}} \Big(H(g_2(z)) + h(g_2(z))\big(L(g_1(z), y) - g_2(z)\big)\Big)\, P(\,\mathrm{d}z,\, \mathrm{d}y) \quad (15)$$

among all increasing functions $g_1 \colon \mathcal{Z} \to \mathbb{R}$ and decreasing $g_2 \colon \mathcal{Z} \to \mathbb{R}$, that is, for $z \preceq z'$ we have $g_1(z) \leq g_1(z')$ and $g_2(z) \geq g_2(z')$. We call any minimizing pair an optimal solution to the isotonic regression problem. Following Jordan, Mühlemann and Ziegel (2020), in order to accommodate the partially ordered set $\mathcal{Z}$, we introduce upper sets $x \subseteq \mathcal{Z}$ to replace single indices $i \in \{1, \ldots, n+1\}$. Set x is said to be an *upper set* if $z \in x$ and $z \preceq z'$ implies $z' \in x$. Let $\mathcal{X}$ consist of all admissible superlevel sets for an increasing function g imposed by the partial order on $\mathcal{Z}$. For a weight vector w of length $\#\mathcal{Z}$ summing to one, we define P^w to be the distribution with density $z \mapsto w(z)$ with respect to P.

As in the case of total orders, keeping either g_1 or g_2 fixed, we can find the optimal solution to (15) with respect to the other component.

Proposition A.1. *(a) Let $g_1 : \mathcal{Z} \to \mathbb{R}$ be given. Then, the optimal antitonic solution $\hat{g}_2$ of (15) corresponding to g_1 is given by*

$$\hat{g}_2(z) = -\min_{x' : z \notin x'} \max_{x \supsetneq x'} -\mathbb{E}(\bar{P}_{x\setminus x'}) = -\max_{x : z \in x} \min_{x' \subsetneq x} -\mathbb{E}(\bar{P}_{x\setminus x'}),$$

where $\bar{P}_{x\setminus x'}$ is the conditional distribution of $L(g_1(Z), Y)$ given $Z \in x \setminus x'$.

(b) Let $g_2 : \mathcal{Z} \to \mathbb{R}$ be given. Then, any optimal isotonic solution $\hat{g}_1$ of (15) with g_2 fixed satisfies

$$\min_{x' : z \notin x'} \max_{x \supsetneq x'} T^-(P^w_{x\setminus x'}) \leq \hat{g}_1(z) \leq \max_{x : z \in x} \min_{x' \subsetneq x} T^+(P^w_{x\setminus x'}),$$

where $P^w_{x\setminus x'}$ is the conditional law of P^w on the event $x \setminus x'$, where the weight w is proportional to $h(g_2(z)), z \in \mathcal{Z}$.

Proof. Follows with the same argument as for total orders. □

As in Section 3.2, we need to introduce some notation for the investigations ahead. In the following, we denote an optimal solution on the subset $\bar{x} \subseteq \mathcal{Z}$ by $\hat{g}_{1;\bar{x}}$ and by $\hat{g}_1\big|_{\bar{x}}$ we denote the optimal solution on the original set restricted to $\bar{x}$.

Thinking in terms of superlevel sets, Lemma 3.3 states that $\hat{g}_1\big|_{\mathcal{Z}\setminus\bar{x}} \leq \hat{g}_{1;\mathcal{Z}\setminus\bar{x}}$ and $\hat{g}_1\big|_{\bar{x}} \geq \hat{g}_{1;\bar{x}}$ for any $\bar{x} \in \mathcal{X}$.

Lemma A.2. *Let $\bar{x} \in \mathcal{X}$ and assume that*

$$\hat{g}_1(z) := \min_{x' : z \notin x'} \max_{x \supsetneq x'} T^\lambda(P^w_{x\setminus x'}) = \max_{x : z \in x} \min_{x' \subsetneq x} T^\lambda(P^w_{x\setminus x'})$$

for some $\lambda \in [0, 1]$. Then we have $\hat{g}_1\big|_{\bar{x}} \geq \hat{g}_{1;\bar{x}}$ and $\hat{g}_1\big|_{\mathcal{Z}\setminus\bar{x}} \leq \hat{g}_{1;\mathcal{Z}\setminus\bar{x}}$.

Proof. It suffices to notice that

$$\hat{g}_1\big|_{\bar{x}}(z) = \max_{x:z\in x}\min_{x'\subsetneq x} T^\lambda(P^w_{x\setminus x'}) \geq \max_{\substack{x\in\mathcal{X};\, x\subseteq\bar{x};\\ z\in x}}\ \min_{x'\in\mathcal{X};\, x'\subsetneq x} T^\lambda(P^w_{x\setminus x'}) = \hat{g}_{1;\bar{x}}(z).$$

The second statement follows with similar reasoning. □

Let us recall the following observations made in Jordan, Mühlemann and Ziegel (2020). For fixed weights w, we can minimize

$$\int_{x\times\mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y), \quad \text{for all } \eta\in\mathbb{R} \tag{16}$$

among all admissible superlevel sets x for an increasing function $g_1 : \mathcal{Z}\to\mathbb{R}$ to obtain an optimal solution to (15). The search for the optimal superlevel set x needs to be conducted for every $\eta\in\mathbb{R}$. Again there is a one-to-one correspondence between admissible superlevel sets and optimal solutions. Instead of an increasing function $\iota : \mathbb{R}\to\{1,\dots,n+1\}$ with $\iota(\eta)\in I_{1:n}(\eta)$ for all η, we now have a monotone function $\xi : \mathbb{R}\to\mathcal{Z}$, in the sense that $\xi(\eta')\subseteq\xi(\eta)$ for $\eta'>\eta$. Moreover, it should hold that $\xi(\eta)\in X_{\mathcal{Z}}(\eta)$ for all $\eta\in\mathbb{R}$, where $X_{\mathcal{Z}}(\eta)\subseteq\mathcal{X}$ denotes the set of all superlevel sets minimizing (16). Then the correspondence between an optimal solution $\hat{g}_1$ and $\xi(\eta)$ is given by

$$\inf\{\eta : z\notin\xi(\eta)\} = \hat{g}_1(z) = \max\{\eta : z\in\xi(\eta)\}.$$

The next result is the generalization of Lemma 3.4 to partial orders.

Lemma A.3. *Let $\bar{x}\in\mathcal{X}$. We have that $X_{\mathcal{Z}}(\eta)\cap\mathcal{X}_{\mathcal{Z}\setminus\bar{x}}\subseteq X_{\mathcal{Z}\setminus\bar{x}}(\eta)$, where $X_{\bar{x}}(\eta)$ is the set of minimizing superlevel sets for the isotonic regression problem (15) for the distribution of (Z, Y) conditional on $Z\in\bar{x}\subseteq\mathcal{Z}$ and $\mathcal{X}_{\mathcal{Z}\setminus\bar{x}}$ is the set of all upper sets in $\mathcal{Z}\setminus\bar{x}$.*

Proof. Let $x'\in X_{\mathcal{Z}}(\eta)\cap\mathcal{X}_{\mathcal{Z}\setminus\bar{x}}$ for some $\eta\in\mathbb{R}$. Therefore, the function

$$t_\eta : \mathcal{X}\to\mathbb{R}, x\mapsto\int_{x\times\mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y)$$

has a minimum at x'. We can write

$$\int_{x\times\mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y) = \int_{x\cap(\mathcal{Z}\setminus\bar{x})\times\mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y) + \int_{x\cap\bar{x}\times\mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y).$$

Hence, $t_\eta\big|_{\mathcal{X}_{\mathcal{Z}\setminus\bar{x}}}$ has a minimum at x' and thus $x'\in X_{\mathcal{Z}\setminus\bar{x}}(\eta)$. If t_η has a minimum in $x=\bar{x}$, then

$$t_\eta(x) - \int_{x\cap\bar{x}\times\mathbb{R}} V(\eta, y)\, P^w(\,\mathrm{d}y) \geq 0,$$

with equality in $x=\bar{x}$. Thus, $\emptyset\in X_{\mathcal{Z}\setminus\bar{x}}(\eta)$. □

Let us generalize Proposition 3.7 to partial orders.

Proposition A.4. *For fixed g_2, corresponding $\hat{g}_1^-$ and any other solution $\hat{g}_1$ to (15) for g_2, we have for any decreasing g_2'*

$$\int_{\mathcal{Z}\times\mathbb{R}} \tilde{L}(g_1^-(z), g_2'(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y) \le \int_{\mathcal{Z}\times\mathbb{R}} \tilde{L}(g_1(z), g_2'(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y).$$

Proof. Let $\mathcal{Q}$ and $\mathcal{Q}^-$ denote the partition of $\mathcal{Z}$ corresponding to $\hat{g}_1$ and $\hat{g}_1^-$, respectively. By Lemma 2.1, it suffices to show that for all $\eta \in \mathbb{R}$

$$\int_{\mathcal{Z}\times\mathbb{R}} S_{\eta,2}(\hat{g}_1^-(z), g_2'(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y) \le \int_{\mathcal{Z}\times\mathbb{R}} S_{\eta,2}(\hat{g}_1(z), g_2'(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y).$$

For the latter, it suffices to show that for all $\bar{x} \in \mathcal{X}$

$$\int_{\bar{x}\times\mathbb{R}} L(\hat{g}_1^-(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y) \le \int_{\bar{x}\times\mathbb{R}} L(\hat{g}_1(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y).$$

Again it suffices to consider $\hat{g}_1$ with superlevel sets in $\cup_\eta X(\eta)$ and again we will prove the converse. In other words, for all $\bar{x} \in \mathcal{X}$, we have

$$\int_{\mathcal{Z}\setminus\bar{x}\times\mathbb{R}} L(\hat{g}_1(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y) \le \int_{\mathcal{Z}\setminus\bar{x}\times\mathbb{R}} L(\hat{g}_1^-(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y). \tag{17}$$

If $\mathcal{Z}\setminus\bar{x} = Q_1 \cup \dots \cup Q_i$, $Q_1, \dots, Q_i \in \mathcal{Q}$, Lemma A.3 implies that $\hat{g}_1\big|_{\mathcal{Z}\setminus\bar{x}}$ is optimal for the distribution of (Z, Y) conditional on $Z \in \mathcal{Z}\setminus\bar{x}$. Thus, (17) holds trivially. If there exists no sequence of partition elements such that $\mathcal{Z}\setminus\bar{x} = Q_1 \cup \dots \cup Q_i$ we distinguish two cases.

Case 1: If $\mathcal{Z}\setminus\bar{x} = Q_1^- \cup \dots \cup Q_{i^-}^-$, $Q_1^-, \dots, Q_{i^-}^- \in \mathcal{Q}^-$ Lemma A.2 implies that

$$\hat{g}_1^-\big|_{\mathcal{Z}\setminus\bar{x}} = \hat{g}_{1;\mathcal{Z}\setminus\bar{x}}^- \le \hat{g}_1\big|_{\mathcal{Z}\setminus\bar{x}} \le \hat{g}_1^+\big|_{\mathcal{Z}\setminus\bar{x}} \le \hat{g}_{1;\mathcal{Z}\setminus\bar{x}}^+$$

Moreover, by Lemma A.3, $X_{\mathcal{Z}}(\eta) \cap \mathcal{X}_{\mathcal{Z}\setminus\bar{x}} \subseteq X_{\mathcal{Z}\setminus\bar{x}}(\eta)$. Hence $\xi\big|_{\mathcal{Z}\setminus\bar{x}}(\eta) \in X_{\mathcal{Z}\setminus\bar{x}}(\eta)$ for all $\eta \in \mathbb{R}$, where $\xi : \mathbb{R} \to \mathcal{Z}$ is the function imposing the score-minimizing superlevel sets corresponding to $\hat{g}_1$. Thus, by Proposition 4.5 in Jordan, Mühlemann and Ziegel (2020) $\hat{g}_1\big|_{\mathcal{Z}\setminus\bar{x}}$ is an optimal solution to the isotonic regression problem for the distribution of (Z, Y) conditional on $Z \in \mathcal{Z}\setminus\bar{x}$.

Case 2: It remains to consider the case where no sequence of partition elements such that $\mathcal{Z}\setminus\bar{x} = Q_1^- \cup \dots \cup Q_{i\_}^-$ exists. Note that $\hat{g}_1$ is optimal for all $z \in \mathcal{Z}\setminus\bar{x}$ with $\hat{g}_{1;\mathcal{Z}\setminus\bar{x}}^-(z) \le \hat{g}_1(z)$. Indeed, for those z, we have $\bar{g}_{1;\mathcal{Z}\setminus\bar{x}}^-(z) \le \hat{g}_1(z) \le \hat{g}_{1;\mathcal{Z}\setminus\bar{x}}^+(z)$, and can argue as in case 1. For $z \in \mathcal{Z}\setminus\bar{x}$ with $\hat{g}_{1;\mathcal{Z}\setminus\bar{x}}^-(z) > \hat{g}_1(z)$, we can argue similarly as in the proof of Proposition 3.7. For every $z \in \{z' \in \mathcal{Z}\setminus\bar{x} : \hat{g}_{1;\mathcal{Z}\setminus\bar{x}}^-(z') > \hat{g}_1(z')\}$ we have $z \in Q_{i+r}$, $r \in \{1, \dots, k\}$. Moreover, $\hat{g}_1$ is constant on each Q_{i+r}, $r \in \{1, \dots, k\}$. With the same reasoning as in the proof of Proposition 3.7, we obtain that

$$\int_{Q_{i+r}^{>}\times\mathbb{R}} L(\hat{g}_{1;\mathcal{Z}\setminus\bar{x}}^-(z), y)\, P(\,\mathrm{d}z,\,\mathrm{d}y) \le \int_{Q_{i+r}^{>}\times\mathbb{R}} L(c_i, y)\, P(\,\mathrm{d}z,\,\mathrm{d}y)$$

$$\leq \int_{Q^{>}_{i+r}\times\mathbb{R}} L(c_i^-, y)\, P(\,\mathrm{d}z,\,\mathrm{d}y)$$

for all $r \in \{1,\ldots,k\}$, where $Q^{>}_{i+r} := Q_{i+r} \cap \{z \in \mathcal{Z} \setminus \bar{x} : \hat{g}^-_{1;\mathcal{Z}\setminus\bar{x}}(z) > \hat{g}_1(z)\}$. This implies the statement. □

Proposition 3.8 also translates directly to partial orders.

Proposition A.5. *Assume that there exist $\hat{g}_1, \hat{g}_2 \colon \mathcal{Z} \to \mathbb{R}$ minimizing (15). Then $\hat{g}_1^-(\cdot; h(\hat{g}_2))$, and the corresponding $\hat{g}_2^-(\cdot; \hat{g}_1^-(\cdot; h(\hat{g}_2)))$ are also minimizers.*

Proof. The argument is the same as in the proof of Proposition 3.8. □

As in the case of total orders, a simultaneously optimal solution may not necessarily exist, since $\hat{g}_2^-$ imposes additional constraints. Nonetheless, we are able to formulate a criterion so that simultaneous optimality is reached whenever the criterion is fulfilled. Let

$$\hat{g}_1(z) = \min_{x' : z \notin x'} \max_{x \supsetneq x'} T^-(P_{x\setminus x'}) = \max_{x : z\in x} \min_{x' \subsetneq x} T^-(P_{x\setminus x'}),$$
$$\hat{g}_2(z) = -\min_{x' : z \notin x'} \max_{x \supsetneq x'} -\mathbb{E}(\bar{P}_{x\setminus x'}) = -\max_{x : z\in x} \min_{x' \subsetneq x} -\mathbb{E}(\bar{P}_{x\setminus x'}),$$

where $P_{x\setminus x'}$ is the conditional law of P on the event $x \setminus x'$.

Proposition A.6. *Let $\hat{g}_1^-$, $\hat{g}_2^-$ as defined above. A simultaneously optimal solution exists if and only if $\hat{g}_1^- = \hat{g}^-_{1;\mathcal{Z}\setminus\bar{x}}$ for all superlevel sets $\mathcal{Z}\setminus\bar{x}$, $\bar{x} \in \mathcal{X}$ assumed by $-\hat{g}_2^-$.*

The reasoning behind this Proposition is analogous to the reasoning behind Proposition 3.9.

Acknowledgments

The authors are grateful for constructive comments of two anonymous reviewers, which improved the paper.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, London. MR0326887

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

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. *The Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **78** 505–562. MR3506792

Dimitriadis, T., Fissler, T. and Ziegel, J. F. (2020). The efficiency gap. Preprint, arXiv:2010.14146.

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *The Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **78** 505–562. MR3506792

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics* **44** 1680–1707. MR3519937

Fissler, T. and Ziegel, J. F. (2019). Order-sensitivity and equivariance of scoring functions. *Electronic Journal of Statistics* **13** 1166–1211. MR3935847

Frongillo, R. and Kash, I. A. (2020). Elicitation Complexity of Statistical Properties. *Biometrika*. In press, https://doi.org/10.1093/biomet/asaa093. MR4341356

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106** 746–762. MR2847988

Henzi, A., Ziegel, J. F. and Gneiting, T. (2019). Isotonic Distributional Regression. arXiv:1909.03725. MR4349124

Jordan, A. I., Mühlemann, A. and Ziegel, J. F. (2020). Optimal solutions to the isotonic regression problem. arXiv:1904.04761v2. MR4417369

Lambert, N. S. (2019). Elicitation and evaluation of statistical forecasts. *Preprint*. Stanford University, Stanford. (Available from http://ai.stanford.edu/~nlambert/papers/elicitability.pdf).

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics* **14** 24–49. MR4047593

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: perspectives for banking regulation. *The Annals of Applied Statistics* **11** 1901–1911. MR3743282

Patton, A. J. (2020). Comparing Possibly Misspecified Forecasts. *Journal of Business & Economic Statistics* **38** 796–809. MR4154889

Patton, A. J., Ziegel, J. F. and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and Value-at-Risk). *Journal of Econometrics* **211** 388–413. MR3980129

Steinwart, I., Pasin, C., Williamson, R. and Zhang, S. (2014). Elicitation and Identification of Properties. In *Proceedings of the 27th Conference on Learning Theory* 1–45.

Ziegel, J. F. (2016). Coherence and elicitability. *Mathematical Finance* **26** 901–918. MR3551510

Ziegel, J. F., Krüger, F., Jordan, A. and Fasciati, F. (2020). Robust forecast evaluation of expected shortall. *Journal of Financial Econometrics* **18** 95–120.