Electronic Journal of Statistics
Vol. 11 (2017) 4488–4514
ISSN: 1935-7524
DOI: 10.1214/17-EJS1365

Bounded isotonic regression

Ronny Luss

IBM Research AI
T.J. Watson Research Center
Yorktown Heights, NY 10549, USA
e-mail: rluss@us.ibm.com

and

Saharon Rosset

Department of Statistics and Operations Research
Tel Aviv University
Tel Aviv, Israel
e-mail: saharon@post.tau.ac.il

Abstract: Isotonic regression offers a flexible modeling approach under monotonicity assumptions, which are natural in many applications. Despite this attractive setting and extensive theoretical research, isotonic regression has enjoyed limited interest in practical modeling primarily due to its tendency to suffer significant overfitting, even in moderate dimension, as the monotonicity constraints do not offer sufficient complexity control. Here we propose to regularize isotonic regression by penalizing or constraining the range of the fitted model (i.e., the difference between the maximal and minimal predictions). We show that the optimal solution to this problem is obtained by constraining the non-penalized isotonic regression model to lie in the required range, and hence can be found easily given this non-penalized solution. This makes our approach applicable to large datasets and to generalized loss functions such as Huber's loss or exponential family log-likelihoods. We also show how the problem can be reformulated as a Lasso problem in a very high dimensional basis of upper sets. Hence, range regularization inherits some of the statistical properties of Lasso, notably its degrees of freedom estimation. We demonstrate the favorable empirical performance of our approach compared to various relevant alternatives.

MSC 2010 subject classifications: Primary 62G08; secondary 62J07.
Keywords and phrases: Multivariate isotonic regression, nonparametric regression, regularization path, range regularization, lasso regularization.

Received February 2017.

Contents

1. Introduction

The statistics community has long had an interest in fitting monotone models to data [18, 2, 29, 25]. Assume we are given a set of observations $(x_1, y_1), ..., (x_n, y_n)$, each consisting of a vector of covariates $x \in \mathcal{X}$ and a response $y \in \mathbb{R}$, along with a partial order $\preceq$ on covariate space $\mathcal{X}$. The class of isotonic models $\mathcal{G}$ is defined as the collection of models on $\mathcal{X}$ which obey the partial order constraints, i.e., a model $g : \mathcal{X} \to \mathbb{R}$ is in $\mathcal{G}$ if for all $x, z \in \mathcal{X}$,

$$x \preceq z \Rightarrow g(x) \leq g(z).$$

The partial order on $\mathcal{X}$ can be translated to a set of order constraints on the observations indexed by $\mathcal{I} = \{(i, j) : x_i \preceq x_j\}$. Isotonic regression aims to find the model $\hat{g} \in \mathcal{G}$ which minimizes the residual sum of squares on the sample data:

$$\hat{g} = \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n} (y_i - g(x_i))^2. \tag{1}$$

It is well-known that an optimal solution of (1) can be written as a non-negative linear combination of upper sets plus an intercept. An upper set is any subset $U \subsetneq \{x_1, ..., x_n\}$ such that $x_i \in U$, $x_i \preceq x_j$ implies $x_j \in U$. Assume we have N distinct upper sets (N is typically exponential in n) denoted $U_1, ..., U_N$. We thus know that the optimal solution has the form

$$\hat{g}(x) = \sum_{l=1}^{N} \hat{\beta}_l \mathbb{I}\{x \in U_l\} + \hat{\alpha}, \tag{2}$$

for some set of nonnegative coefficients $\hat{\beta}_l \geq 0$. Note that this only defines the solution at our observed covariate vectors $x_1, \ldots, x_n$, but it can easily be extended to all of $\mathcal{X}$ by associating every point $x \in \mathcal{X}$ with the set of upper sets it "dominates". Prediction and extrapolation in this context have been previously discussed in Luss et al. [21].

The most commonly used covariate space and partial order are the standard Euclidean space $\mathcal{X} = \mathbb{R}^d$ and the partial order defined on $x, z \in \mathbb{R}^d$ by $x \preceq z$ iff the inequality holds coordinate-wise, i.e., $x_j \leq z_j$ for all $j = 1, \ldots, d$.

In terms of model form, isotonic regression is clearly very attractive in situations where monotonicity is a reasonable assumption but commonly assumed structures such as linearity or additivity are not. Indeed, this formulation has found useful applications in biology [25, 21], medicine [29], statistics [2] and psychology [18], among others (see Tibshirani et al. [31] for a nearly-isotonic formulation that relaxes the isotonic assumption). However, two major concerns arise when considering the practical use of isotonic regression in *modern* situations as the number of observations n, the data dimensionality d, and the number of isotonicity constraints $m = |\{(i, j) : x_i \preceq x_j\}|$ implied by (1) all grow large: statistical overfitting and computational difficulty. The notations n, d, and m will refer to these quantities throughout the paper.

The first concern is statistical overfitting. Beyond very low dimensions, the isotonicity constraints on the family $\mathcal{G}$ can become inefficient in controlling model complexity and the isotonic regression solutions can be severely overfitted (for example, see Bacchetti [1] and Schell and Singh [29]). At the extreme, there may be no isotonicity constraints because no two observations obey the coordinate-wise requirement for the partial order $\preceq$. In this case, every observation is a singleton upper set, and if we denote these n upper sets by $U_1, ..., U_n$, the isotonic model can fit the data perfectly using only these upper sets:

$$\hat{g}(x) = \sum_{l=1}^{n} (y_l - \min_i y_i) \mathbb{I}\{x \in U_l\} + \min_i y_i,$$

i.e., $\alpha = \min_i y_i$ and $\hat{\beta}_l = y_l - \min_i y_i$ for all $l = 1, \ldots, n$, providing a perfect interpolation of the training data. As demonstrated in the literature [29, 21] and below, the overfitting concern is clearly well-founded when considering the optimal isotonic regression model implied by (1), even in non-extreme cases with a large number of constraints. In this case, regularization, i.e. fitting isotonic models that are constrained to a restricted subset of $\mathcal{G}$, offers an approach that maintains isotonicity while controlling variance, leading to improved accuracy.

A second important issue is computation. Traditional methods developed in the statistics community did not scale well with the dimension d [19, 3]. However, in recent years, ideas from optimization have permeated this area and led to the development of very efficient algorithms which can be used to solve problems with tens of thousands of observations in any dimension. The most efficient algorithm for isotonic regression known to the authors is due to Hochbaum and Queyranne [14] where the problem is cast in a more general form they refer to as the convex cost closure problem. They show how to obtain the global isotonic solution with the complexity of solving a single minimum-cut problem (which deals with finding a minimal cut through the arcs of a graph). Furthermore, their algorithm can impose fixed upper and lower bounds on the isotonic model at all observation points.

The isotonic recursive partitioning (IRP) method proposed in Spouge et al. [30] and Luss et al. [21] (following previous related work by Maxwell and Muckstadt [23] and Roundy [28], among others) finds a solution of (1) through iterative partitioning of the space $\mathcal{X}$ (i.e., solving a sequence of minimum-cut problems). Each split can be computed efficiently and the procedure is guaranteed to converge to an optimal solution of (1). It offers a highly efficient approach for solving (1) that is also amenable to regularization through early stopping of the iterative partitioning process. However, as demonstrated there, IRP does not offer sufficient complexity control and regularization in many cases. For example, at dimension $d = 6$ for particular simulation models, the first iteration of IRP already performs more than half of the fitting of the optimal solution of (1) as measured by equivalent degrees of freedom [21]. IRP regularization also lacks a rigorous formulation of the problem solved, as one cannot explicitly write the optimization problem being solved to obtain the model after k IRP iterations.

We are thus interested in designing a rigorous regularization approach for isotonic regression that would allow for a continuum of regularized models with increasing model complexity. In the nonparametric context where we fit a model by choosing it from a large function class, there has been an extensive use of smoothness (or complexity) penalties, leading to important tools such as spline methods and kernel machines [12]. These nonparametric smoothness regularization problems are typically solved by identifying "equivalent" parametric representations and solving those using ridge- or lasso-like approaches. This includes smoothing splines, kernel machines, and total variation penalties [22], among others. Importantly, total variation penalties are intimately tied to Lasso-type penalties on spline bases, as demonstrated by Mammen and van der Geer [22] and others.

A similar approach can be proposed for isotonic regression, where a natural measure of model "complexity" is the *range* of model predictions. For a model $g \in \mathcal{G}$, define:

$$\operatorname{range}(g) = \sup_{x \in \mathcal{X}} g(x) - \inf_{x \in \mathcal{X}} g(x).$$

It is easy to see that for a model of the form (2), we have:

$$\operatorname{range}(\hat{g}) \leq \sum_l \hat{\beta}_l, \tag{3}$$

so there is a close connection here too between range regularization and Lasso regularization, which we return to later.

Our explicit range-regularized isotonic regression model is thus (in its constrained form):

$$\hat{g} = \operatorname*{argmin}_{g \in \mathcal{G}, \mathsf{range}(g) \leq s} \sum_i (y_i - g(x_i))^2, \tag{4}$$

or in its equivalent penalized form:

$$\hat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \sum_i (y_i - g(x_i))^2 + \lambda \cdot \operatorname{range}(g). \tag{5}$$

Our main observation in this paper is that this problem has a simple optimal solution, obtained by solving the non-regularized isotonic regression problem and finding optimal thresholds on this solution which obey the range constraint. Thus, all solutions to range-regularized isotonic regression problems are in fact thresholded versions of the non-regularized solutions. This leads to a simple and efficient algorithm for deriving all range-regularized solutions, which we term Bounded Isotonic Regression (BIR) and present in Section 2.

In Section 3 we investigate the properties of our new formulation and algorithm. We show that BIR is in fact equivalent to solving a non-negative Lasso problem in the set of upper sets. We further examine the regularization behavior of BIR and demonstrate that it generally adds one degree of freedom per upper-set added to the solution, as shown for Lasso by Zou et al. [33]. Hence, BIR offers a *gentle* fitting of isotonic models with increasing complexity. Finally, we show that the basic ideas and efficient implementation of the BIR framework are not limited to isotonic regression, but can be applied to isotonic modeling problems with other differentiable loss functions (e.g., exponential family log-likelihoods or Huber's loss).

We demonstrate our BIR algorithm on simulated datasets in Section 4. We compare BIR regularization to IRP, and show its continuous regularization behavior, as well as its improved predictive performance, compared to IRP. We demonstrate its superior predictive performance over multivarate additive regression splines (MARS) for models that are both isotonic and highly complex. We also compare model selection approaches for the regularization parameter in BIR: general purpose cross-validation is compared to methods based on in-sample error, such as AIC and GCV, using the degrees of freedom approximation based on the Lasso connection.

It is interesting to compare our approach to a recent paper by Fang and Meinshausen [9] which also combines isotonic regression and Lasso penalties in an algorithm termed LISO (Lasso-Isotone). Fang and Meinshausen limit their interest to *additive* isotonic models, i.e., where a univariate isotonic function is fit to every covariate, and the overall isotonic model is the sum of these univariate functions. Additivity significantly limits the generality of the isotonic models being fit, but allows Fang and Meinshausen [9] to fit useful models to very high dimensional data (large d). Our approach is more assumption-free, and in lower dimensions (up to $d = 8$ or $d = 10$) when the models are complex and data is abundant, demonstrates superior performance relative to LISO in our experiments (Section 4).

A corresponding result to our main observation about a simple solution for range-regularized isotonic regression problem (5) was independently derived in parallel for the range-constrained isotonic regression problem (4) in Chen et al. [6]. While our results were discovered independently, our proof proceeds in similar fashion. However, whereas our result on the range-regularized case directly leads to an algorithm for the entire path of solutions (i.e., a solution to problem (5) for all values of λ), the result in Chen et al. [6] for the range-constrained case was used to prove a theorem about the degrees of freedom of the estimator. Again, we have independently discovered the same properties, however

our claims were realized through the above-mentioned connection between BIR and Lasso. We address the connections to Chen et al. [6] more thoroughly in Sections 2.1 and 3.1.1

2. Regularization path for bounded isotonic regression

In this section, we derive an efficient algorithm for solving BIR. It relies on the result in Theorem 1 (in Section 2.1 below) that the solutions to BIR are all thresholded versions of the non-regularized isotonic regression solution.

2.1. Derivation of the optimal solution to BIR

Our focus is on the range-regularized isotonic regression problem in its penalized form given by problem (5). We first reformulate it as

$$\begin{array}{ll} \min_{\hat{y},\hat{a},\hat{b}} & \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda(\hat{b}-\hat{a}) \\ \text{subject to} & \hat{y}_i \le \hat{y}_j \text{ for all } (i,j)\in\mathcal{I} \\ & \hat{a}\le \hat{y}_i \le \hat{b} \text{ for all } i\in\{1,\dots,n\}, \end{array} \tag{6}$$

where $\hat{a}=\inf\{\hat{y}_i : i\in\{1,\dots,n\}\}$, $\hat{b}=\sup\{\hat{y}_i : i\in\{1,\dots,n\}\}$, and the range of the isotonic model is now captured by $\hat{b}-\hat{a}$. We will use the optimality conditions to (6) to derive an algorithm for efficiently generating the solution to (6) for all values of λ. The optimality conditions are:

(a) $2(\hat{y}_i-y_i)+\sum_{j:(i,j)\in\mathcal{I}}\mu_{ij}-\sum_{j:(j,i)\in\mathcal{I}}\mu_{ji}-\gamma_i+\delta_i=0$ for all $i\in\{1,\dots,n\}$
(b) $\hat{y}_i\le\hat{y}_j$ for all $(i,j)\in\mathcal{I}$
(c) $\hat{a}\le\hat{y}_i\le\hat{b}$ for all $i\in\{1,\dots,n\}$
(d) $\gamma^T 1=\lambda$, $\delta^T 1=\lambda$
(e) $\mu_{ij}(\hat{y}_i-\hat{y}_j)=0$ for all $(i,j)\in\mathcal{I}$
(f) $\gamma_i(\hat{a}-\hat{y}_i)=0, \delta_i(\hat{b}-\hat{y}_i)=0$ for all $i\in\{1,\dots,n\}$
(g) $\gamma,\delta,\mu\ge 0$,

where $\mu\in\mathbf{R}^{|\mathcal{I}|}$, $\gamma\in\mathbf{R}^n$, $\delta\in\mathbf{R}^n$ are the dual variables to the monotonicity constraints, the lower range, and the upper range constraints, respectively.

The following theorem shows that BIR solutions are all thresholded versions of the non-regularized isotonic regression solution (i.e., BIR with $\lambda=0$). Thus, if we have this non-regularized solution, we can obtain the solution to all BIR problems with minimal effort. We note that this result extends a similar theorem of Fang and Meinshausen [9] from one dimensional (complete order) isotonic regression to partial order isotonic regression. Furthermore, a corresponding result was recently made in Chen et al. [6] for solving the constrained version of bounded isotonic regression. We discuss their result in more detail at the end of this section.

Theorem 1. *Let $\hat{z}$ be the optimal solution to the non-regularized isotonic regression problem. Then setting*

$$\hat{y}_i = \max\left(\hat{a}, \min\left(\hat{z}_i, \hat{b}\right)\right) \tag{7}$$

for all $i \in \{1, \ldots, n\}$ solves BIR problem (6), where $\hat{a}$ and $\hat{b}$ solve the equations

$$2\sum_i \left(\hat{a} - \hat{z}_i\right)_+ = \lambda \qquad \text{and} \qquad 2\sum_i \left(\hat{z}_i - \hat{b}\right)_+ = \lambda.$$

Proof. By construction, we prove that $\hat{y}$ as defined by (7) solves the optimality conditions (a)–(f) for problem (6) given above.

Let μ^* be the optimal dual variables to the corresponding monotonicity constraints in the non-regularized problem (which has the same optimality conditions above when $\lambda = 0$). Then $\lambda = 0$ implies $\gamma = \delta = 0$ and condition (a) can be rewritten as

$$\hat{z}_i = y_i - \frac{1}{2}\left(\sum_{j:(i,j)\in\mathcal{I}} \mu^*_{ij} - \sum_{j:(j,i)\in\mathcal{I}} \mu^*_{ji}\right) \text{ for all } i \in \{1, \ldots, n\} \tag{8}$$

For $\lambda > 0$, set the dual variables to be the same as the optimal dual variables as given in (8), i.e., set $\mu = \mu^*$. First note that the optimality conditions (e), called complementarity conditions, are satisfied by construction ($\mu_{ij} > 0 \Rightarrow \hat{z}_i = \hat{z}_j \Rightarrow \hat{y}_i = \hat{y}_j$ and $\hat{y}_i < \hat{y}_j \Rightarrow \hat{z}_i < \hat{z}_j \Rightarrow \mu_{ij} = 0$). The next set of complementarity conditions (f) imply that either γ_i or δ_i can be nonzero, but not both. Nonnegativity of γ and δ, along with condition (a) which can be written $2(\hat{y}_i - \hat{z}_i) = \gamma_i - \delta_i$, then imply $\gamma_i = 2(\hat{y}_i - \hat{z}_i)_+$ and $\delta_i = 2(\hat{z}_i - \hat{y}_i)_+$. Given the definition for $\hat{y}$ in (7), note that $(\hat{y}_i - \hat{z}_i)_+ = (\max(\hat{a} - \hat{z}_i, \min(0, \hat{b} - \hat{z}_i)))_+ = (\hat{a} - \hat{z}_i)_+$ and $(z^*_i - \hat{y}_i)_+ = (\min(\hat{z}_i - \hat{a}, \max(0, \hat{z}_i - \hat{b})))_+ = (\hat{z}_i - \hat{b})_+$, so that $\gamma_i = 2(\hat{a} - \hat{z}_i)_+$ and $\delta_i = 2(\hat{z}_i - \hat{b})_+$.

Next are the primal variables. Suppose $\hat{y}_i = \max(\hat{a}, \min(\hat{z}_i, \hat{b}))$ for some $\hat{a}, \hat{b}$. With this definition, conditions (d) are equivalent to $2\sum_i (\hat{a} - \hat{z}_i)_+ = \lambda$ and $2\sum_i (\hat{z}_i - \hat{b})_+ = \lambda$. Both $\hat{a}$ and $\hat{b}$ are scalars and can be chosen to satisfy these equations, and then used to obtain $\hat{y}$ from its definition. Given optimal $\hat{y}$, the optimal dual variables can be computed.

Optimality conditions (a),(d), and (g) hold by construction. Condition (b) holds from the definition of $\hat{y}$ since $\hat{z}$ is feasible for the non-regularized problem, and condition (c) holds strictly by definition of $\hat{y}$. Conditions (f) hold by construction and using condition (c): $\hat{y}^*_i \neq \hat{a} \Rightarrow \hat{y}^*_i > \hat{a} \Rightarrow \hat{z}^*_i > \hat{a} \Rightarrow \gamma_i = 2(\hat{y}_i - \hat{z}_i)_+ = 2(\max(\hat{a} - \hat{z}_i, \min(0, \hat{b} - \hat{z}_i)))_+ = 0$, since both terms in the max are less than or equal to zero. A similar argument shows the other complementarity condition holds as well. Optimality conditions (a)–(g) hold and, thus, $\hat{y}, \hat{a}, \hat{b}$ are optimal solutions to the range-regularized isotonic regression problem (6). □

Theorem 1 is illustrated in Figure 1 where the solution to BIR problem (6) for a decreasing sequence of λ values is depicted. We use data from a well-known

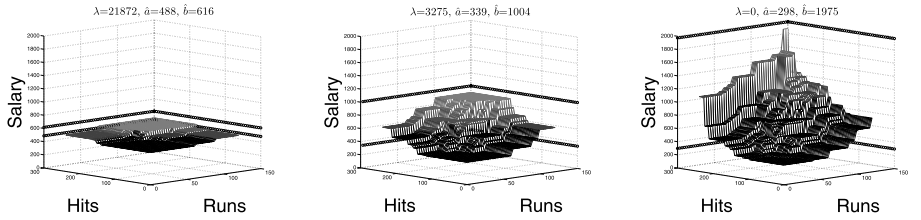

FIG 1. *Illustration of BIR on Baseball data. Salary is modeled by number of runs batted in and hits. BIR models corresponding to a decreasing sequence of* λ *concluding at the non-regularized model* ($\lambda = 0$) *are shown. The optimal* $\hat{a}, \hat{b}$ *described in Theorem 1 are given by the bottom and top lines in each figure, respectively. They correspond to the range for which the non-regularized solution in the last figure is thresholded (i.e., the first two figures with* $\lambda > 0$ *are thresholded versions of the last figure* ($\lambda = 0$) *with thresholds determined by Theorem 1).*

Baseball dataset [13] which describes the dependence of salary on a collection of player properties. Models are limited to two covariates in order to facilitate visualization. The number of runs batted in and hits were selected since they seemed a-priori most likely to comply with the isotonicity assumptions. The increasing range, given in each figure as the distance between the two lines corresponding to the optimal values of $\hat{a}$ and $\hat{b}$, can be seen as λ is decreased. The first two figures with $\lambda > 0$ are thresholded versions of the last figure which depicts the non-regularized solution, i.e. the surface between the $\hat{a}$ and $\hat{b}$ lines is identical to the corresponding surface in the non-regularized solution.

As mentioned above, Chen et al. [6] independently and concurrently developed a corresponding thresholding theory for the constrained version of bounded isotonic regression. In our notation, we could reformulate problem (4) as

$$\begin{array}{ll} \min_{\hat{y},\hat{a}} & \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ \text{subject to} & \hat{y}_i \le \hat{y}_j \text{ for all } (i,j) \in \mathcal{I} \\ & \hat{a} \le \hat{y}_i \le \hat{a} + s \text{ for all } i \in \{1,\dots,n\} \end{array} \tag{9}$$

where, again, $\hat{a} = \inf\{\hat{y}_i : i \in \{1,\dots,n\}\}$, $\hat{a} + s = \sup\{\hat{y}_i : i \in \{1,\dots,n\}\}$, and the range is still captured by s. Rather than adding variables $\hat{a}$ with our range constraints, Chen et al. [6] constrains the range by adding constraints of the form $y_j - y_i \le s$ for all (i,j) such that there exists no observations that bound y_i from below and no observations that bound y_j from above (with respect to the partial order). We next state the main result of Proposition 3.3 of Chen et al. [6] using our notation and in similar form to Theorem 1 above:

Proposition 3.3 of Chen et al. [6]. *Let* $\hat{z}$ *be the optimal solution to the non-regularized isotonic regression problem. Then setting*

$$\hat{y}_i = \max(\hat{a}, \min(\hat{z}_i, \hat{a} + s)) \tag{10}$$

for all $i \in \{1,\dots,n\}$ *solves range-constrained BIR problem (9), where* $\hat{a}$ *solves the equation*

$$\sum_i (\hat{a} - \hat{z}_i)_+ - \sum_i (\hat{z}_i - \hat{a} - s)_+ = 0.$$

The optimality conditions for problem (9) are almost equivalent to those of problem (6) where the only differences are

(c) $\hat{a} \leq \hat{y}_i \leq \hat{a} + s$ for all $i \in \{1, \ldots, n\}$
(d) $\gamma^T 1 - \delta^T 1 = 0$
(e) $\gamma_i(\hat{a} - \hat{y}_i) = 0, \delta_i(\hat{a} + s - \hat{y}_i) = 0$ for all $i \in \{1, \ldots, n\}$

Given the optimality conditions, proof of the above proposition follows straightforward almost exactly as the proof for Theorem 1. The most significant difference is that the new equation for determining $\hat{a}$ is given by the new optimality condition (d). We note that this proof differs from that given in Chen et al. [6] where they describe the optimality conditions in terms of a graph representation of problem (9) that falls under a class of problems known as transportation problems. Proposition 3.3 of Chen et al. [6] also states that $\hat{a}$ is a non-increasing function of s (which we did not mention in Theorem 1, but this is trivially noted from the equation that determines $\hat{a}$ above in Proposition 3.3 of Chen et al. [6]). Chen et al. [6] developed the above theory in order to prove a statement about the degrees of freedom of the BIR estimator, and we discuss this further in Section 3.1.1. Rather, our motivation of such theory is to derive efficient algorithms for generating solutions to problem (6) for a sequence of λ.

Lastly, regarding the above discussion on thresholding isotonic regression, an interesting and related work was done by Hu [15]. While he previously showed that problem (6) with fixed $\hat{a}$ and $\hat{b}$ could be solved by projecting the non-regularized isotonic regression solution onto the bound constraints, he showed that the same was not true when the lower (or upper) bounds for each data point are not equivalent. The new approach he took approximated the bounded isotonic regression problem with an extended and weighted isotonic regression problem and showed that the limit (as some of the weights go to infinity) of solving these weighted isotonic regression problems converges to the bounded isotonic regression problem (with fixed but different bounds).

2.2. Regularization path

Given Theorem 1, we can compute BIR for any fixed value of the regularization parameter λ. Rather than compute solutions over a grid of λ which would require solving the one variable equations for $\hat{a}$ and $\hat{b}$, we next derive an efficient path algorithm that generates all solutions for a sequence of λ. For fixed λ, define the following three sets of indices: $\mathcal{A}^\lambda = \{i : \hat{a} < \hat{y}_i < \hat{b}\}$, $\mathcal{A}_a^\lambda = \{i : \hat{a} = \hat{y}_i\}$, $\mathcal{A}_b^\lambda = \{i : \hat{b} = \hat{y}_i\}$. The decreasing sequence of λ values for which we derive solutions contains those values at which the sets $\mathcal{A}^\lambda$, $\mathcal{A}_a^\lambda$, and $\mathcal{A}_b^\lambda$ change. Between those values the interpolation is trivial.

Our approach to problem (6) requires that we first solve the non-regularized isotonic regression problem (1) to get $\hat{z}$, which can be done using the IRP algorithm of [21]. We next consider the solution to the completely regularized problem when the range is constrained to zero, i.e., $\lambda = \infty$ and $\hat{b} = \hat{a}$. Under

this constraint, $\hat{y}_i$ is fixed to the same constant for all i and it is trivial to see that $\hat{y}_i = \overline{y} = \hat{a} = \hat{b}$. Given this fit, we can obtain the dual variables γ and δ and determine that $\lambda = \gamma^T 1$. It is easy to see that we would obtain the same solution for any larger value of λ.

Our next goal is to decrease λ until there is a change in $\mathcal{A}^\lambda$, $\mathcal{A}_a^\lambda$, or $\mathcal{A}_b^\lambda$, in which case $\hat{a}$ must be decreased and $\hat{b}$ must be increased. Clearly, from optimality condition (f), $\gamma_i = \delta_i = 0$ for all $i \in \mathcal{A}^\lambda$. From the above conditions $2(\hat{y}_i - \hat{z}_i) = 2(\hat{a} - \hat{z}_i) = \gamma_i$ for all $i \in \mathcal{A}_a^\lambda$ and $2(\hat{y}_i - \hat{z}_i) = 2(\hat{b} - \hat{z}_i) = -\delta_i$ for all $i \in \mathcal{A}_b^\lambda$, we see that $\hat{a}$ is decreased by decreasing γ_i for all $i \in \mathcal{A}_a^\lambda$ and $\hat{b}$ is increased by decreasing δ_i for all $i \in \mathcal{A}_b^\lambda$. We make two observations: for each $i \in \mathcal{A}_a^\lambda$, γ_i must be decreased by the same amount in order for the optimality condition to be maintained, and similarly, for each $i \in \mathcal{A}_b^\lambda$, δ_i must be decreased by the same amount. A change in the sets will occur when a nonzero γ_i or δ_i becomes zero.

Let $\gamma_{\min} = \min\{\gamma_i : i \in \mathcal{A}_a^\lambda\}$ and $\delta_{\min} = \min\{\delta_i : i \in \mathcal{A}_b^\lambda\}$ be the most that either dual variable can be decreased. Recall that decreasing the dual variables decreases λ. Decreasing γ or δ as much as possible would decrease λ by $\gamma_{\min}|\mathcal{A}_a^\lambda|$ and $\delta_{\min}|\mathcal{A}_b^\lambda|$, respectively. If $\gamma_{\min}|\mathcal{A}_a^\lambda| < \delta_{\min}|\mathcal{A}_b^\lambda|$, we will decrease γ as much as possible (by $\gamma_{\min}$ for all $i \in \mathcal{A}_A^\lambda$) and decrease $\hat{a}$ by $\gamma_{\min}/2$ and λ by $\gamma_{\min}|\mathcal{A}_a^\lambda|$. Note that δ cannot be decreased by as much as possible because then optimality condition (d) would not hold. Summing over the optimality condition for $\hat{b}$ over $i \in \mathcal{A}_b^\lambda$ and using optimality condition (d) gives $\hat{b} = (\sum_{i \in \mathcal{A}_b^\lambda} z_i^* - \lambda/2)/|\mathcal{A}_b^\lambda|$. The new δ is obtained from our formula $\delta_i = (\hat{z}_i - \hat{b})_+$. Given the new boundaries $\hat{a}$ and $\hat{b}$, the new model $\hat{y}$ for the new regularization λ can be computed. The cases $\gamma_{\min}|\mathcal{A}_a^\lambda| > \delta_{\min}|\mathcal{A}_b^\lambda|$ and $\gamma_{\min}|\mathcal{A}_a^\lambda| = \delta_{\min}|\mathcal{A}_b^\lambda|$ are handled in a similar manner. Obtaining the BIR regularization path is summarized by Algorithm 1. The output is the path of $\hat{a}$ and $\hat{b}$, which can be used to generate the path of isotonic models via $\hat{y}_i = \max(\hat{a}, \min(\hat{z}_i, \hat{b}))$.

While we do not develop a corresponding algorithm for problem (9) for a sequence of s, we note that such an algorithm could be a subject of future research. Initial investigations suggest that formulation (9) may not lead to a simple algorithm because it is difficult to control how parameter s and variable a move together. Another reformulation where we add variable $\hat{b}$, change the upper bound constraints to $\hat{y}_i \le \hat{b}$, and control the range with the constraint $\hat{b} - \hat{a} \le s$ might lead to a corresponding efficient algorithm.

3. Properties of bounded isotonic regression

Here we discuss the statistical behavior of our new algorithm and the resulting models. We address two aspects:

1. The connection between BIR and Lasso, and the resulting conclusions about regularization behavior of the BIR "regularization path", using results on degrees of freedom of Lasso [33].
2. Generalization of our algorithm to other loss functions besides the quadratic loss in (5), which turns out to be straightforward.

Algorithm 1 Bounded Isotonic Regression (BIR)

Require: Optimal nonregularized isotonic regression solution $\hat{z}$ and mean of observations $\overline{y}$

Initialize primal variables $\hat{y}_j^{(0)} = \hat{a}^{(0)} = \hat{b}^{(0)} = \overline{y}$
Initialize dual variables $\gamma_i^{(0)} = (\overline{y} - \hat{z}_i)_+$, $\delta_i^{(0)} = (\hat{z}_i - \overline{y})_+$
Initialize $\lambda^{(0)} = \sum_i \gamma_i^{(0)}$
$i = 0$
while $\hat{a}^{(i)} > \min_j \hat{z}_j$ and $\hat{b}^{(i)} < \max_j \hat{z}_j$ **do**
  Let $\mathcal{A}_a^\lambda = \{j : \hat{a}^{(i)} = \hat{y}_j^{(i)}\}$ and $\mathcal{A}_b^\lambda = \{j : \hat{b}^{(i)} = \hat{y}_j^{(i)}\}$
  Let $\gamma_{\min} = \min\{\gamma_j : j \in \mathcal{A}_a^\lambda\}$ and $\delta_{\min} = \min\{\delta_j : j \in \mathcal{A}_b^\lambda\}$
  if $\gamma_{\min}|\mathcal{A}_a^\lambda| < \delta_{\min}|\mathcal{A}_b^\lambda|$ **then**
    $\hat{a}^{(i+1)} = \hat{a}^{(i)} - \gamma_{\min}$
    $\gamma_j^{(i+1)} = (\hat{a}^{(i+1)} - \hat{z}_j)_+$ for all j
    $\lambda^{(i+1)} = \sum_j \gamma_j^{(i+1)}$ for all j
    $\hat{b}^{(i+1)} = (\sum_{j \in \mathcal{A}_b^\lambda} z_j^* - \lambda^{(i+1)}/2)/|\mathcal{A}_b^\lambda|$
    $\delta_j^{(i+1)} = (\hat{z}_j - \hat{b}^{(i+1)})_+$ for all j
  else if $\gamma_{\min}|\mathcal{A}_a^\lambda| > \delta_{\min}|\mathcal{A}_b^\lambda|$ **then**
    $\hat{b}^{(i+1)} = \hat{b}^{(i)} + \delta_{\min}$
    $\delta_j^{(i+1)} = (\hat{z}_j - \hat{b}^{(i+1)})_+$ for all j
    $\lambda^{(i+1)} = \sum_j \delta_j^{(i+1)}$ for all j
    $\hat{a}^{(i+1)} = (\sum_{j \in \mathcal{A}_a^\lambda} z_j^* + \lambda^{(i+1)}/2)/|\mathcal{A}_a^\lambda|$
    $\gamma_j^{(i+1)} = (\hat{a}^{(i+1)} - \hat{z}_j)_+$ for all j
  else
    $\hat{a}^{(i+1)} = \hat{a}^{(i)} - \gamma_{\min}$
    $\hat{b}^{(i+1)} = \hat{b}^{(i)} + \delta_{\min}$
    $\gamma_j^{(i+1)} = (\hat{a}^{(i+1)} - \hat{z}_j)_+$ for all j
    $\delta_j^{(i+1)} = (\hat{z}_j - \hat{b}^{(i+1)})_+$ for all j
    $\lambda^{(i+1)} = \sum_j \gamma_j^{(i+1)}$ for all j
  end if
  $\hat{y}_j^{(i+1)} = \max(\hat{a}^{(i+1)}, \min(\hat{z}_j, \hat{b}^{(i+1)}))$
  $i = i + 1$
end while
return $\hat{a}^{(i)}$ and $\hat{b}^{(i)}$ for all iterations

3.1. BIR as a Lasso problem

As noted in Section 1, any isotonic fit to data can be described as a non-negative linear combination of N upper set indicators, and the range of any such model is smaller or equal than the sum of upper set coefficients, as shown in (3). Here we show that in fact any optimal solution to (5) is also exactly an optimal solution to a non-negative Lasso problem in upper sets. Following (2), we propose the following Lasso-like problem, with added non-negativity requirement:

$$(\hat{\beta}, \hat{\alpha}) = \arg\min_{\beta \geq 0, \alpha} \sum_{i=1}^{n} (y_i - \sum_{l=1}^{N} \beta_l \mathbb{I}\{x_i \in U_l\} - \alpha)^2 + \lambda \sum_{l=1}^{N} \beta_l, \tag{11}$$

where we optimize a penalized (penalty on the sum of upper set coefficients)

criterion over all isotonic functions. Denote the optimal solution:

$$\hat{g}_\lambda(x) = \sum_l \hat{\beta}_l \mathbb{I}\{x \in U_l\} + \hat{\alpha}.$$

Lemma 2. *The upper sets for which $\hat{\beta}_l > 0$ in the solution to (11) are a nested sequence, i.e., if $\hat{\beta}_{l_1} > 0$ and $\hat{\beta}_{l_2} > 0$ then either $U_{l_1} \subset U_{l_2}$ or $U_{l_2} \subset U_{l_1}$.*

Proof. Assume by negation that this does not hold for l_1, l_2. $U_u = U_{l_1} \cup U_{l_2}$ and $U_{12} = U_{l_1} \setminus U_{l_2}$ and $U_{21} = U_{l_2} \setminus U_{l_1}$ are all also trivially upper sets (U_{21} or U_{12} may be empty). Now we increase $\hat{\beta}_{U_u}$ by $\min(\hat{\beta}_{l_1}, \hat{\beta}_{l_2})$, increase $\hat{\beta}_{U_{12}}$ by $\hat{\beta}_{l_1} - \min(\hat{\beta}_{l_1}, \hat{\beta}_{l_2})$ and increase $\hat{\beta}_{U_{21}}$ by $\hat{\beta}_{l_2} - \min(\hat{\beta}_{l_1}, \hat{\beta}_{l_2})$, and set $\hat{\beta}_{l_1} = \hat{\beta}_{l_2} = 0$. The new $\hat{g}(x)$ remained unchanged for all x, but $\sum_l \hat{\beta}$ has decreased by $\min(\hat{\beta}_{l_1}, \hat{\beta}_{l_2}) > 0$. Hence this new function would have a lower penalized loss than the original $\hat{g}_\lambda$ in (11), which contradicts the optimality of $\hat{g}_\lambda$. □

Lemma 3.

$$\mathit{range}(\hat{g}_\lambda) = \sum_l \hat{\beta}_l$$

Proof. By Lemma 2 the upper sets for which $\hat{\beta}_l > 0$ are a nested sequence. Let $l_{\min}$ be the index of the minimal set in this sequence. Then for some $\tilde{x} \in l_{\min}$, the fitted value is

$$\hat{g}_\lambda(\tilde{x}) = \sum_l \hat{\beta}_l + \hat{\alpha},$$

because this $\tilde{x}$ is in all nested sequence. On the other hand, let $l_{\max}$ be the index of the maximal set in this sequence. Then $U_{l_{\max}} \subsetneq \mathcal{X}$ because if $U_{l_{\max}} = \mathcal{X}$, then we can set $\hat{\beta}_{l_{\max}} = 0$ and increase the non-penalized $\hat{\alpha}$ correspondingly, again contradicting optimality. Hence there is $x^* \notin U_{l_{\max}}$ such that $\hat{g}_\lambda(x^*) = \hat{\alpha}$. Hence

$$\text{range}(\hat{g}_\lambda) \geq \hat{g}_\lambda(\tilde{x}) - \hat{g}_\lambda(x^*) = \sum_l \hat{\beta}_l.$$

The other inequality is trivial, and has been stated in the introduction. □

Theorem 4. *Any optimal solution to the Lasso problem (11) is also an optimal solution to the BIR problem (5) with the same λ.*

Proof. From Lemma 3, the optimal solution to (11) also gives a solution to (5) with the same penalized objective value. Denote this solution by $\hat{g}_\lambda$, as before. It remains to prove that (5) cannot have a better solution.

Assume by negation there is a lower penalized objective solution to (5), denoted by $\hat{f}_\lambda$. Assume WLOG that $\{\hat{f}_\lambda(x_i) : i = 1, ..., n\}$ is sorted in increasing order, and denote $U_i^* = \{x : \hat{f}_\lambda(x) \geq \hat{f}_\lambda(x_i)\}$. It is easy to verify that $U_1^*, ..., U_n^*$ are a nested sequence of upper sets, and that we can express:

$$\hat{f}_\lambda(x) = \sum_{i=2}^{n} (\hat{f}_\lambda(x_i) - \hat{f}_\lambda(x_{i-1})) \mathbb{I}\{x \in U_i^*\} + \hat{f}_\lambda(x_1),$$

which is a solution for (11) with the same loss and penalty as $\hat{f}_\lambda$ has in (5), hence same penalized objective. This contradicts optimality of $\hat{g}_\lambda$. □

3.1.1. Implications of the Lasso connection

Lasso-type problems have been extensively studied in the literature from various theoretical, methodological and practical perspectives. We focus here on two important classes of Lasso-related results and their implications on BIR.

First is the statistical complexity of BIR models, as measured in degrees of freedom (df) and optimism [7, 12]. For a generic penalized Lasso problem,

$$\hat{\beta}(\lambda) = \arg\min \sum_i (y_i - x_i^T \hat{\beta})^2 + \lambda \sum_j |\beta_j|,$$

denote by $\mathcal{A} = \{j : \hat{\beta}_j(\lambda) \neq 0\}$ the set of active covariates with non-zero coefficients. Zou et al. [33] have shown that Stein's unbiased estimator for df is the number of covariates with non-zero coefficients in the solution:

$$\hat{df} = |\mathcal{A}|.$$

If our Lasso-like formulation for BIR (11) did not contain the non-negativity constraint, this result would apply directly and would imply, using Lemma 2, that the number of blocks (distinct values of the function $\hat{g}_\lambda$) is the Stein estimate.

In the presence of the non-negativity constraint, we claim that the generic Lasso result still holds. This can be verified by carefully considering the proof of Zou et al. [33], which only relies on the behavior of the Lasso regularization path *given* the set of active variables. Since the non-negativity constraint only affects selection of variables into the active set, and not the path direction given this active set, the result still holds. Furthermore, considering Algorithm 1, it is easy to see that the BIR solutions only add variables into the model and never drop them. Hence the Stein unbiased estimate of degrees of freedom is simply the number of iterations the algorithm has performed. This result is also consistent with the result of Meyer and Woodroofe [24], showing that the Stein estimate of degrees of freedom of non-regularized isotonic regression is the number of blocks (groups of distinct values) in the solution, which is the number of iterations the algorithm takes to reach the solution as $\lambda \to 0$.

The Lasso connection is not the only way to make our claim about the degrees of freedom of the BIR estimator. Indeed, Chen et al. [6] (equation (22)), again independently and concurrently, show the result that the number of degrees of freedom is equal to the number of blocks in the BIR solution. While they derive the result for range-constrained BIR, the claim is equivalent although they pose it in terms of the number of connected components of a graph induced by the BIR estimator. It is interesting to note two completely different approaches for obtaining this result. We rely on results for the Lasso regularization path, while Chen et al. [6] rely on a result from algebraic graph theory pertaining to the

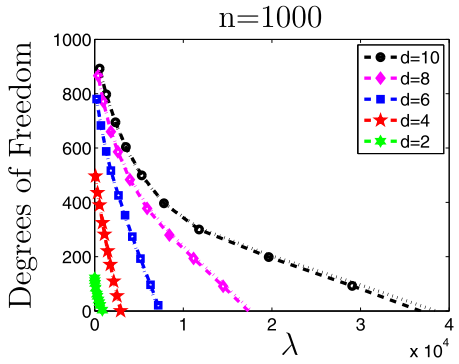

FIG 2. *Degrees of freedom vs BIR iterations. The X axis is the value of the regularization parameter* λ*. The Y axis shows the degrees of freedom as estimated from repeated simulations. For each simulation setup, the dashed lines with the denoted shapes show the actual* df *at each regularization level (estimated directly using the optimism theorem), while the dotted line (in most cases obscured by the dashed line) shows the empirical mean of* $\hat{df}$ *(number of pieces in the lasso solution, identical to the number of iterations of Algorithm 1) at each regularization level. Simulations use i.i.d. covariates* $\mathbf{x}_{ij} \sim \mathcal{U}[0,3]$ *with* $y_i = \prod_j x_{ij} + \mathcal{N}(0, d^2)$*. Each path is the mean over 500 trials with 1000 observations.*

rank of incidence matrices of graphs. Also note that essentially the same result we used for the regularized version Lasso above is also known for the constrained version of Lasso (e.g., see Kato [16]).

Behavior of our BIR regularization path is demonstrated empirically in Figure 2, where we compare the expected value of this Stein estimate to the actual *df* as estimated from repeated simulations using the optimism theorem of Efron [7]. In all dimensions, the number of iterations is seen empirically to be an unbiased estimate of the actual *df*.

This *df* result clarifies the regularization behavior of BIR, and in particular the gradual increase of model complexity as λ decreases. It also naturally facilitates using for BIR the multitude of model selection approaches developed for Lasso. We note that this non-increasing behavior of *df* as λ increases that we illustrate empirically was proved in Chen et al. [6] (refer to their Theorem 3.4 though note there *df* is non-decreasing in λ because λ is the range parameter to the range-constrained case). Our Lasso connection can easily be used to make the same claim since, as we noted above, BIR solutions only add variables and never drop them.

The second aspect we consider is the computational one. Efficient algorithms have been developed for calculating Lasso regularization paths [26, 8]. Can these be used to efficiently calculate BIR solutions? If we consider these algorithms in their standard forms the answer is clearly no, since they include a step of enumerating over all covariates and solving a simple linear equation for each one, which is repeated many times (in every iteration). Since the covariates in our Lasso formulation are upper set indicators, and the number of upper sets is exponential in the number of observations, a direct application of these algo-

rithms is unlikely to yield efficient algorithms. An alternative approach for high dimensional problems could be to replace this enumeration with an appropriate search algorithm as proposed in Rosset et al. [27]. This turns out to be possible for BIR (details are eliminated for brevity), but the resulting algorithm is still substantially less efficient than our algorithmic solution presented in Section 2, which relies on the specific structure of this problem. Hence we do not see a computational benefit in considering the Lasso connection, although this is open for further research.

3.2. Generalization of BIR for other loss functions

A natural question that arises is whether the useful structure identified in Section 2 is unique to isotonic regression with l_2 loss, or whether it generalizes to other loss functions. Of particular interest are exponential family log-likelihoods and robust regression loss functions, as discussed in Luss and Rosset [20] and references therein. The generic problem we consider is:

$$\begin{array}{ll} \min_{\hat{y},\hat{a},\hat{b}} & \sum_{i=1}^{n} L_i(\hat{y}_i) + \lambda(\hat{b}-\hat{a}) \\ \text{subject to} & \hat{y}_i \le \hat{y}_j \text{ for all } (i,j) \in \mathcal{I} \\ & \hat{a} \le \hat{y}_i \le \hat{b} \text{ for all } i \in \{1,\dots,n\}, \end{array} \tag{12}$$

where L_i is some convex and differentiable loss function (usually a function of observation y_i) like the ones mentioned above.

As noted in the Introduction, the non-regularized isotonic modeling problem can be solved for *all* such loss functions with the same complexity as the standard l_2 isotonic regression problem as shown by Hochbaum and Queyranne [14], which addresses the general loss function (and even furthermore solves problem (12) for fixed $\hat{a}$ and $\hat{b}$). Hence for us to be able to solve the BIR version we only need to verify that Theorem 1 also generalizes successfully. This is in fact true, as captured by the following generalized result which is proven in the Appendix.

Theorem 5. *Let $\hat{z}$ be the optimal solution to the non-regularized isotonic regression problem with convex and differentiable loss functions L_i. Then setting*

$$\hat{y}_i = \max\left(\hat{a}, \min\left(\hat{z}_i, \hat{b}\right)\right) \tag{13}$$

for all $i \in \{1,\dots,n\}$ solves the generalized BIR problem (12), where $\hat{a}$ and $\hat{b}$ solve the equations

$$\sum_i \left(\frac{\partial L_i(\hat{a})}{\partial \hat{y}_i} - \frac{\partial L_i(\hat{z}_i^*)}{\partial \hat{y}_i} \right)_+ = \lambda \qquad \text{and} \qquad \sum_i \left(\frac{\partial L_i(\hat{z}_i^*)}{\partial \hat{y}_i} - \frac{\partial L_i(\hat{b})}{\partial \hat{y}_i} \right)_+ = \lambda. \tag{14}$$

Hence, Algorithm 1 can be generalized to solving the family of problems (12) by replacing averages and residuals with their appropriate loss-function-specific generalization, following the same ideas as in Luss and Rosset [20]. We eliminate the details for brevity. We also note that there have been other works regarding bounded isotonic regression with a total order and fixed bounds (see Chakravarti [4] for an example with an l_1 loss function).

4. Experiments

Our simulations examine the performance of BIR from several perspectives: In Section 4.1, we compare BIR and IRP to the popular flexible modeling approach multivariate adaptive regression splines (MARS) [11] as a representative of modern competitive approaches which do not assume linearity or additivity. Reassuringly, BIR shows significantly improved performance for isotonic functions. In Section 4.2, we compare BIR and IRP to additive LISO [9], demonstrating the expected behavior: in lower dimension, with large amounts of data, the added flexibility of BIR allows fitting of better models. In Section 4.3, we concentrate on comparing our two isotonic modeling approaches, IRP [21] and BIR, demonstrating the advantage of BIR regularization as expressed by improved prediction performance. Finally, in Section 4.4, we describe the application of in-sample model selection approaches GCV and AIC to BIR, capitalizing on the Lasso connection, and compare their performance to cross-validation.

Simulations are carried out in the following manner: Training and testing data are generated. The training data is further split into training and validation folds. An IRP model is trained on the training fold resulting in a regularization path. Models along this regularization path are used to generate a path of RMSEs by predicting responses in the validation fold. The model that generates the lowest RMSE is the chosen model which is used to predict the independent testing data. Results on this independent testing data are reported. The global isotonic solution generated by IRP from the training fold of the training data is then used to generate a path of BIR models, which is, in turn, used to select a model using the validation fold of the training data. As with IRP, this chosen BIR model is used to obtain prediction results on the independent testing data. Note that time results for BIR include the time to find the non-regularized solution with IRP. In all simulations, the extra computation that BIR required given the non-regularized solution was less than 5% of the computation required to find the solution. All results are averaged over 50 trials.

4.1. Comparing with MARS

MARS is a well-known regression approach that builds models of the form

$$\hat{f}(x) = \beta_0 + \sum_i \beta_i h_i(x),$$

where each $h_i(x)$ is a hinge function (i.e., of the form $h_i(x) = \max(x_j - t, 0)$ or $h_i(x) = \max(t - x_j, 0)$ for the j^{th} covariate and knot $t \in \mathbf{R}$) or the product of hinge functions. The choice of knots are typically determined by the training data, and β is then estimated by standard linear regression. Basis functions $h_i(x)$ are added to the model in a greedy fashion. Fang and Meinshausen [9] show scenarios where LISO outperforms MARS, particularly in much higher dimensions than are considered here. In lower dimensions, MARS performs very well, but in our experiments suffers computationally with the large number of

TABLE 1

Performance of IRP, BIR, and MARS on two different interaction models with 8 dimensions. Paths of IRP and BIR models are trained on 12000 observations, model selection done on another 3000 observations, and performance results are based on an independent test set of 3000 observations. IRP RMSE *and* BIR RMSE *refer to using models that gave minimum root mean squared errors (RMSE) in model selection to predict the independent test set. Values for RMSE are given along with a conservative 95% confidence interval. Time is measured in seconds. Note that MARS is trained with 15000 observations.*

Model	IRP RMSE	BIR RMSE	MARS RMSE	IRP Time	BIR Time	MARS Time*
1	13.03(± 0.22)	**12.35(± 0.19)**	13.79(± 0.62)	127.9	130.5	228.2
2	14.35(± 0.11)	**14.02(± 0.11)**	14.93(± 0.42)	100.6	102.8	266.1

training observations we consider. In experiments with relatively simple true models and especially low number of observations, MARS and BIR generally perform comparably (results not shown). We concentrate here on situations that are both statistically and computationally challenging, where BIR's advantages are emphasized. Table 1 displays the regression results on two models that incorporate interactions of the variables. The models are

$$\text{Model 1:} \quad z_{ij} \in \{0,1\}, x_{ij} \sim \mathcal{U}[0,5], y_i = \sum_{j=2}^{4} z_{i,j-1} z_{i,j} x_{i,j-1} x_{i,j}^2 + \mathcal{N}(0,5^2)$$

$$\text{Model 2:} \quad z_{ij} \in \{0,1\}, x_{ij} \sim \mathcal{U}[0,5], y_i = \sum_{j=2}^{4} z_{i,j-1} z_{i,j} 2^{x_{i,j-1}} x_{i,j} + \mathcal{N}(0,5^2),$$

where the z_{ij} indicator variables are uniform over $\{0,1\}$ and both z_{ij} and x_{ij} variables are independent. Each model has four indicator variables and four continuous variables for a total of eight dimensions. These are deemed "interaction" models because the terms in the summations are only included in the response if the two corresponding indicators are both on (i.e., both set to one). These are scenarios in which IRP outperforms MARS, as well as where range regularization learns better isotonic models than offered by the regularization of IRP. Furthermore, the timing results clearly show that MARS takes much longer to learn models than IRP and BIR. As the amount of training data is increased, the computational complexity of MARS increases faster than IRP and BIR. The following simulation trains IRP and BIR models on 12000 observations, perform model selection using another 3000 observations, and the performance results are based on an independent test set of 3000 observations. MARS, which has an internal cross-validation mechanism, trains models on the first two folds (15000 observations) with peformance results based on the same independent test set of 3000 observations.

We next use Model 1 to examine how performance and regularization change as the training sample size increases. Figure 3 (left) demonstrates that RMSE on the hold-out test samples decreases as the number of training samples increases, due to better modeling of the entire space. Regarding the regularization parameter λ, the optimal λ increases as the number of training samples increases (not shown). To understand why, consider the BIR formulation as a Lasso-like problem (3). As the number of training samples increases, the number of upper sets (and variables) increases exponentially, and hence a higher regularization

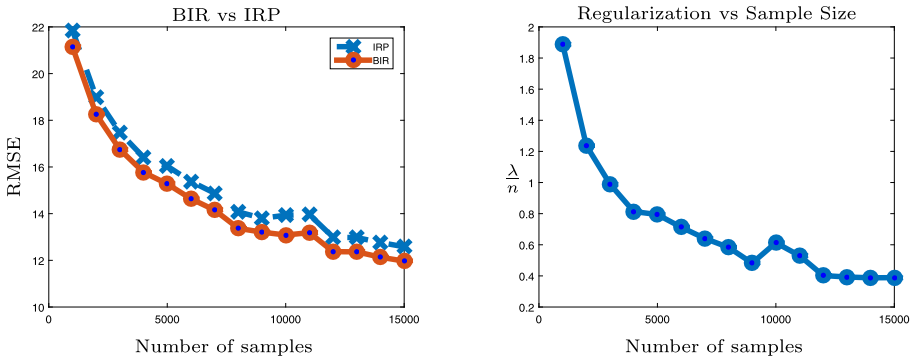

FIG 3. *Regularization as a function of the number of samples for Model 1:* $z_{ij} \in \{0,1\}, x_{ij} \sim \mathcal{U}[0,5], y_i = \sum_{j=2}^{4} z_{i,j-1} z_{i,j} x_{i,j-1} x_{i,j}^2 + \mathcal{N}(0,5^2)$. *The figure on the left demonstrates that accuracy of the model increases (error decreases) as the number of training samples increases. For BIR, the regularization parameter* λ *is optimized over the path for each individual experiment. The figure on the right demonstrates that* $\lambda_n/n \to \lambda_0 \geq 0$. *The number of training samples in both figures varies from 1000 to 15000. Model selection is done on a separate 3000 observations and performance evaluated on another 3000 observations.*

is need to eliminate additional potential upper sets. More importantly, Figure 3 (right) compares the rate of increase in regularization versus the increase in the number of training samples and empirically shows that $\lambda_n = o(n)$ where we denote by λ_n the optimal regularization parameter for n training samples. For a fixed number of variables (which we do not have), this would imply consistency of the BIR model (see Theorem 1 of [17]). Results in our case where the number of variables is exponential in the number of training samples require a Strong Irrepresentable Condition [32]) on the design matrix. Under other technical assumptions, consistency is shown when $\lambda_n \propto n^{\frac{1+c}{2}}$ for some small c (Theorem 4 of [32]). However, while Figure 3 (right) shows that λ_n grows slower than n, our empirical tests do not demonstrate that λ_n grows faster than $\sqrt{n}$ as required for this particular consistency theorem.

4.2. Comparing with LISO

LISO builds models of the form

$$\hat{f}(x) = \alpha + \sum_i h_i(x_i),$$

where $h_i(x_i)$ is a one-dimensional isotonic function of the i^{th} covariate. This additive isotonic regression is trained by taking each $h_i(x_i)$ to be a positive linear combination of the upper sets formed by the i^{th} dimension and solving the following lasso problem,

$$\begin{aligned} \min_{\alpha,\beta} \quad & \sum_{i=1}^{n}(y_i - \alpha - \sum_{j=1}^{d}\sum_{l=1}^{N}\beta_{jl}\mathbb{I}\{x_{ij} \in U_{jl}\})^2 + \lambda \sum_{j=1}^{d}\sum_{l=1}^{N}\beta_{jl} \\ \text{s.t.} \quad & \beta_{jl} \geq 0 \quad \forall j = 1,\ldots,d, l = 1,\ldots,N, \end{aligned} \tag{15}$$

where $\beta \in \mathbf{R}^{d\times N}$. The coefficient for upper set l in the j^{th} dimension (represented by U_{jl}) is β_{jl} and α is again the intercept. Then we have $h_j(x_j) = \sum_{l=1}^{N}\beta_{jl}\mathbb{I}\{x_j \in U_{jl}\}$. Note that the number of upper sets N in each dimension is bounded by the number of training instances n. Then problem (15) reduces to a lasso problem with nd parameters, as shown in Fang and Meinshausen [9], which can be solved by the classic LARS algorithm for low-dimensional problems to get a full regularization path of additive isotonic models. For high-dimensional problems, Fang and Meinshausen [9] offer an algorithm that solves (15) for fixed λ, and hence require solving the problem many times over a λ-grid in order to generate a path of solutions. The following results use a LARS-implementation of LISO in order to obtain a full regularization path for additive isotonic models.

A path of models with increasing complexity (i.e., increasing number of upper sets in the solution) is learned for IRP, BIR, and LISO and performance is shown in Table 2. For dimensions 2-5, the following simulation trains models on 3000 observations, performs model selection using another 3000 observations, and the performance results are based on an independent test set of 3000 observations. The limited training on 3000 observations is because LISO is computationally expensive (3000 observations with d dimensions translates to running LARS on 3000 observations with roughly $3000d$ variables). For dimension 5, we give results for training IRP and BIR with 12000 observations (indicated by 5^*) as well to show how performance improves when training with more data.

Neither model is additive so we should expect isotonic regression (IRP and BIR) to outperform LISO in all dimensions. This is true for the first model, however the second model with $d = 5$ shows better performance for LISO. This is because isotonic regression is less structured than additive isotonic regression and requires more data to learn the model as the dimension increases, demonstrated by the improved performance of IRP and BIR when trained with 12000, rather than 3000, observations (LISO is too computationally expensive to train with this many observations and dimensions as demonstrated by the time results). These simulations further show that range regularization using BIR gives, in many cases, statistically significant improved results over the regularization provided by IRP.

4.3. Comparing with IRP

The previous subsection noted that IRP and BIR are less structured than other regression methods, such as LISO, and hence require more training data to learn a good model. This was exhibited in Table 2 in 5 dimensions by the increased performance between training with 3000 versus 12000 observations. In this section, we give two more examples in slightly higher dimension that are trained with 12000 observations. Results are shown in Table 3. Again, a validation set of 3000 observations is used to select models and performance is measured on

TABLE 2

Performance of IRP, BIR, and LISO on two different isotonic models. Paths of IRP, BIR, and LISO models are trained on 3000 observations, model selection done on another 3000 observations, and performance results are based on an independent test set of 3000 observations. IRP RMSE, BIR RMSE, *and* LISO RMSE *refer to using models that gave minimum root mean squared errors (RMSE) in model selection to evaluate the independent test set. Values for RMSE are given along with a conservative 95% confidence interval. Time is measured in seconds. Note that for dimension 5, results are given for training IRP and BIR with 12000 observations (indicated by* 5^**) to show how performance improves when training with more data.*

Model 1: $y_i = \prod_j x_{ij} + \mathcal{N}(0, 10^2)$ with $x_{ij} \sim \mathcal{U}[0,5]$ and x_{ij} independent						
Dim. (d)	IRP RMSE	BIR RMSE	LISO RMSE	IRP Time	BIR Time	LISO Time
2	**10.12(± 0.04)**	**10.07(± 0.04)**	10.24(± 0.04)	7.5	7.5	37.4
3	10.92(± 0.04)	**10.81(± 0.04)**	13.94(± 0.04)	20.6	20.7	124.6
4	25.83(± 0.31)	**24.49(± 0.25)**	37.40(± 0.26)	86.7	87.1	500.1
5	121.20(± 1.20)	**115.94(± 1.23)**	123.96(± 1.20)	79.3	79.6	478.3
5^*	95.11(± 1.36)	**88.52(± 1.36)**	—	238.3	248.4	—
Model 2: $y_i = \sqrt{\prod_j 2^{x_{ij}}} + \mathcal{N}(0, 5^2)$ with $x_{ij} \sim \mathcal{U}[0,5]$ and x_{ij} independent						
Dim. (d)	IRP RMSE	BIR RMSE	LISO RMSE	IRP Time	BIR Time	LISO Time
2	**5.07(± 0.02)**	**5.05(± 0.02)**	5.30(± 0.03)	6.3	6.3	41.3
3	**5.13(± 0.72)**	**5.02(± 0.70)**	7.93(± 1.11)	32.4	32.5	185.8
4	**10.72(± 2.42)**	**10.14(± 2.17)**	14.40(± 2.52)	55.7	55.9	314.3
5	141.95(± 2.76)	136.81(± 2.78)	**119.85(± 2.84)**	78.0	78.3	527.5
5^*	114.03(± 2.56)	**106.04(± 2.56)**	—	288.8	297.6	—

a separate test set of 3000 observations. Both models in these simulations are highly nonlinear and non-additive. Performance shows that Model 1 is easier to learn than Model 2, but similar trends are seen in the table. BIR improves upon the performance of IRP in dimensions 4, 6, and 8, however, BIR does not improve upon IRP in 10 dimensions. In 10 dimensions, 12000 observations is already insufficient for learning any useful model, since the number of isotonicity constraints becomes too small.

Comparison of performance is further demonstrated in Figure 4, which illustrates performance throughout the regularization path for both IRP and BIR on a single sample of data. Diamonds represent the minimum RMSE along the respective paths. First note that IRP performance is minimized very early in all regularization paths. This is consistent with the observation in Luss et al. [21] that IRP performs most of its fitting in its first few iterations. Next note that the overall trend is the same for IRP and BIR. In dimensions 4,6, and 8, performance improves (RMSE decreases) as the models become more complex until a certain point at which more complexity hurts performance and RMSE increases. The minimum of the BIR path is lower than the minimum of the IRP path for these dimensions under both models, a result due to the sounder and slower regularization employed in BIR compared to early stopping of IRP. This slower model fitting can also be observed by comparing the expected degrees of freedom in Figure 2 to corresponding simulations for IRP in Luss et al. [21].

While each iteration here increases the degrees of freedom by about one, most of the total fitting in IRP was done in the first few iterations (usually more than half in the first iteration). Finally, we again note the equivalent performance in 10 dimensions. Figure 4 depicts the immediate overfitting, attributed to the insufficient constraints, that occurs in both IRP and BIR.

TABLE 3

Performance of IRP and BIR on two different isotonic models. Paths of IRP and BIR models are trained on 12000 observations, model selection done on another 3000 observations, and performance results are based on an independent test set of 3000 observations. IRP RMSE *and* BIR RMSE *refer to using models that gave minimum root mean squared errors (RMSE) in model selection to evaluate the independent test set. Values for RMSE are given along with a conservative 95% confidence interval. Time is measured in seconds.*

Model 1: $y_i = (\sum_{j=1}^{d/2}(x_{ij} + x_{i,j+(d/2)})^2)/d + \mathcal{N}(0, 30^2)$ with $x_{ij} \sim \mathcal{U}[0, 10]$ and x_{ij} independent				
Dim. (d)	IRP RMSE	BIR RMSE	IRP Time	BIR Time
4	31.43(± 0.11)	**30.94(± 0.11)**	97.3	98.9
6	32.63(± 0.15)	**31.55(± 0.13)**	281.7	288.2
8	35.81(± 0.15)	**33.84(± 0.13)**	214.3	220.4
10	35.48(± 0.13)	35.43(± 0.13)	444.6	455.4
Model 2: $y_i = (\sum_{j=3}^{d}(x_{i,j-1} + x_{i,j-2})^{x_{i,j}})/d + \mathcal{N}(0, 50^2)$ with $x_{ij} \sim \mathcal{U}[0, 10]$ and x_{ij} independent				
Dim. (d)	IRP RMSE	BIR RMSE	IRP Time	BIR Time
4	132.90(± 2.19)	**125.05(± 2.00)**	297.2	304.6
6	276.08(± 2.95)	**263.98(± 2.88)**	400.8	410.3
8	405.67(± 2.72)	**389.71(± 2.60)**	428.0	438.2
10	383.37(± 2.96)	383.37(± 2.96)	515.3	529.8

4.4. Comparing model selection approaches for BIR

Our experiments so far have exclusively used cross-validation (CV) for selecting the regularization parameters of BIR and other approaches. This is convenient for comparing between different approaches which may not necessarily have available in-sample approaches for model selection. CV is also an appropriate model selection approach for observational situations where future prediction points are independently drawn ("random-X"), which arguably represent the majority of modern data analysis scenarios [12].

However, for BIR specifically, the Lasso connection allows us to use a variety of in-sample model selection methods developed or adapted for Lasso, including AIC, GCV and others [10, and references therein]. Here we briefly compare CV to AIC and GCV as approaches for selecting λ in BIR, where we use the Stein estimate *df* as the number of parameters / degrees of freedom in AIC and GCV.

Because the in-sample approaches are targeted at "fixed-X" situations, we slightly change the setup of the simulation for this experiment, and draw the test set for evaluating selected models at the same set of x-values as the training

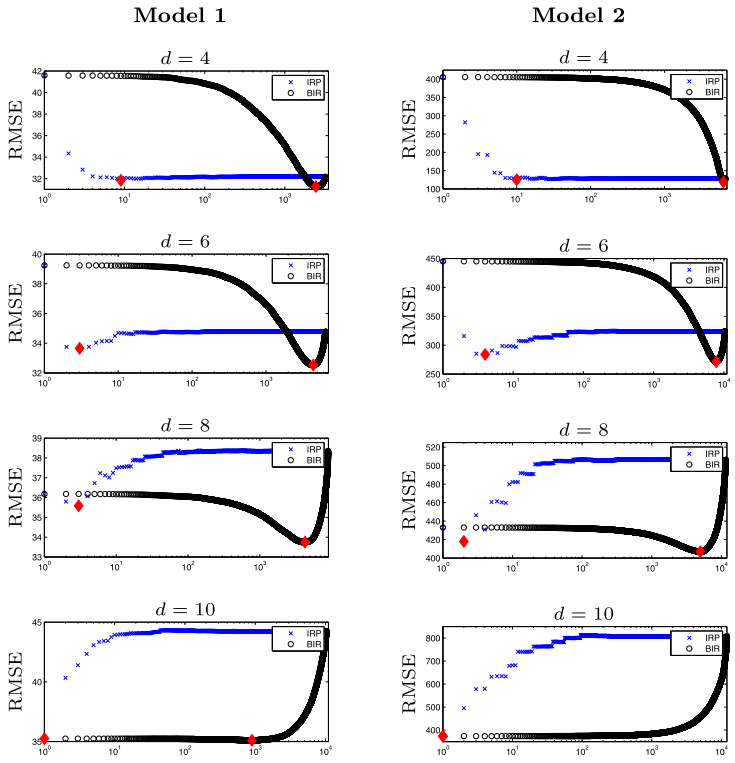

FIG 4. *Root mean squared error (RMSE) for out-of-sample predictions of simulations with different dimensions d. The x-axis in each figure corresponds to the number of iterations made by IRP and BIR in log-scale, i.e. the curves show how the RMSE of test data varies as IRP and BIR progress. Diamonds indicate minimum RMSE along the paths. Models are those used in Table 3.*

set. As before, for CV a portion of the training set is set aside for model selection, while for AIC and GCV the entire set (15000 observations) is used for model building. In general, in this situation, we expect in-sample approaches to do better than CV.

We compare the approaches on models 1 and 2 from Section 4.3, and show the results in Table 4. For model 1, which is a relatively simple model with limited non-additivity effects, CV does roughly as well as GCV and AIC and essentially selects the same models. Still, GCV appears to be slightly superior to both CV and AIC in higher dimensions. Model 2 is a much more "wild" model, hence prediction at new covariate vectors outside the training set is a much more difficult task than "fixed-X" prediction. Consequently for this model the difference between the CV model selection and GCV/AIC becomes evident. CV selects models with heavy regularization, while GCV/AIC tend to select smaller λ and hence models which are appropriate for predicting under the "fixed-X" assumption. The exception is the deteriorated performance of GCV

TABLE 4

Performance of model selection approaches for λ in BIR on two different isotonic models. For CV, BIR models are trained on 12000 observations, and model selection is done on another 3000 observations, while for GCV/AIC the models are trained on all 15000 observations. Performance results are based on an independent test set of 3000 observations, whose covariate vectors are resampled from the training set to represent the "fixed-X" situation. Values for RMSE are given along with a conservative 95% confidence interval

Model 1: $y_i = (\sum_{j=1}^{d/2}(x_{ij} + x_{i,j+(d/2)})^2)/d + \mathcal{N}(0, 30^2)$ with $x_{ij} \sim \mathcal{U}[0, 10]$ and x_{ij} independent			
Number of Variables (d)	CV RMSE	GCV RMSE	AIC RMSE
4	31.52(± 0.10)	31.55(± 0.10)	31.59(± 0.10)
6	33.07(± 0.14)	33.33(± 0.13)	33.88(± 0.12)
8	33.97(± 0.15)	33.92(± 0.15)	36.15(± 0.18)
10	34.87(± 0.10)	33.83(± 0.08)	38.47(± 0.17)
Model 2: $y_i = (\sum_{j=3}^{d}(x_{i,j-1} + x_{i,j-2})^{x_{i,j}})/d + \mathcal{N}(0, 50^2)$ with $x_{ij} \sim \mathcal{U}[0, 10]$ and x_{ij} independent			
Number of Variables (d)	CV RMSE	GCV RMSE	AIC RMSE
4	93.14(± 3.24)	60.28(± 0.23)	60.28(± 0.23)
6	177.68(± 3.44)	67.17(± 0.25)	67.17(± 0.25)
8	304.78(± 4.21)	69.82(± 0.28)	69.82(± 0.28)
10	383.37(± 2.82)	294.35(± 37.55)	70.57(± 0.23)

at dimension $d = 10$, a phenomenon whose detailed investigation may reveal further insights but is a topic for future study.

We also performed some experiments comparing the performance of the three approaches in prediction in the "random-X" scenario. As expected, CV was consistently superior to AIC/GCV in this setting (results not shown).

5. Discussion and conclusion

In this paper we propose to regularize isotonic regression by penalizing or constraining the range of the estimated function and name this new approach BIR. We demonstrate that, given the non-regularized isotonic regression model, all BIR solutions can be generated by a simple and efficient algorithm because they are obtained by thresholding the non-regularized solution from above and below. Furthermore, the BIR problem can be formulated as a non-negative Lasso problem in the basis of upper set indicators and thus inherits properties of Lasso, in particular its regularization behavior. Like Lasso, it adds about one degree of freedom with each iteration (each upper set added to the model). Thus, BIR combines a sound regularization approach, efficient computations, and interesting connections to other methods. Furthermore, we show that the BIR algorithm can easily be generalized to other loss functions, including exponential family log-likelihoods and robust regression. This significantly enhances the utility of BIR methods.

As mentioned in the Introduction, isotonic regression suffers from overfitting issues that severely limit its utility for modern high-dimensional problems. Our

simulations demonstrate that BIR can significantly increase the range of usefulness of isotonic modeling compared to the non-regularized solution or coarsely regularized alternative when using IRP. Up to about dimension eight, BIR can significantly improve on isotonic regression and still generate useful models. Overall, we believe our simulations demonstrate that when isotonicity assumptions are appropriate, the true relationship is complex and non-additive, the dimension is relatively low, and data is abundant, properly regularized isotonic regression is likely to do very well compared to alternatives. It should be noted that our simulations are "space filling" in the sense that the covariate values are uniformly distributed in the covariate space $\mathcal{X}$. This means that the actual dimension is also the effective dimension. Natural data are often highly structured (as captured for example by PCA) and can be closely approximated by lower dimensional spaces. In our context it means that the isotonic constraints required to control model complexity can persist into higher dimension in natural data than in our simulations, thus allowing BIR to remain useful.

An interesting connection of BIR is to total variation penalties, which have become important in several application domains [22, 5]. In $d = 1$ dimension, the range of a monotonic function is trivially also its total variation, so BIR can be thought of as a total variation approach with added isotonicity constraints, rather than isotonic regression with an added range constraint. In higher dimensions, total variation definitions become mathematically quite involved, but for isotonic functions they simply reduce back to the range. Hence BIR can also be thought of as total variation penalized isotonic regression.

Appendix

Proof of Theorem 5

As with the proof for Theorem 1, we prove that the solution given in the theorem solves the optimality conditions. The optimality conditions for the generalized BIR problem (12) are the same as for BIR problem (6) with the generalized first-order optimality condition:

$$\text{(a):}\quad \frac{\partial L_i(\hat{y}_i)}{\partial \hat{y}_i} + \sum_{j:(i,j)\in\mathcal{I}} \mu_{ij} - \sum_{j:(j,i)\in\mathcal{I}} \mu_{ji} - \gamma_i + \delta_i = 0 \text{ for all } i \in \{1,\dots,n\}.$$

We assume that $\hat{a} < \hat{b}$ (the theorem trivially holds in equality). Let μ^* be the optimal dual variables to the corresponding monotonicity constraints in the non-regularized problem (which has the same optimality conditions above when $\lambda = 0$). Then $\lambda = 0$ implies $\gamma = \delta = 0$ and condition (a) can be rewritten as

$$\frac{\partial L_i(\hat{z}_i)}{\partial \hat{y}_i} = -\sum_{j:(i,j)\in\mathcal{I}} \mu^*_{ij} + \sum_{j:(j,i)\in\mathcal{I}} \mu^*_{ji} \text{ for all } i \in \{1,\dots,n\}.$$

Take the dual variables μ to be equal to the dual variables μ^* of the non-regularized problem. Optimality condition (e) again holds immediately by construction. Optimality conditions (f) imply that either γ_i or δ_i can be nonzero,

but not both. Nonnegativity of γ and δ, along with condition (a) which can be written

$$\frac{\partial L_i(\hat{y}_i)}{\partial \hat{y}_i} - \frac{\partial L_i(\hat{z}_i)}{\partial \hat{y}_i} = \gamma_i - \delta_i$$

then imply

$$\gamma_i = \left(\frac{\partial L_i(\hat{y}_i)}{\partial \hat{y}_i} - \frac{\partial L_i(\hat{z}_i)}{\partial \hat{y}_i} \right)_+ \quad \text{and} \quad \delta_i = \left(\frac{\partial L_i(\hat{z}_i)}{\partial \hat{y}_i} - \frac{\partial L_i(\hat{y}_i)}{\partial \hat{y}_i} \right)_+.$$

Suppose $\hat{y}$ is defined by (13). Then

$$\begin{aligned} \hat{y}_i^* \neq \hat{a} \Rightarrow \hat{y}_i^* > \hat{a} \Rightarrow \hat{z}_i^* > \hat{a} \Rightarrow \gamma_i &= \left(\tfrac{\partial L_i(\hat{y}_i)}{\partial \hat{y}_i} - \tfrac{\partial L_i(\hat{z}_i)}{\partial \hat{y}_i} \right)_+ \\ &= \left(\tfrac{\partial L_i(\max(\hat{a}, \min(\hat{z}_i, \hat{b})))}{\partial \hat{y}_i} - \tfrac{\partial L_i(\hat{z}_i)}{\partial \hat{y}_i} \right)_+ = 0, \end{aligned}$$

where equality to zero is by convexity of $L_i(\cdot)$ and since $\hat{z}_i^* > \hat{a} \Rightarrow \hat{z}_i^* \geq \max(\hat{a}, \min(\hat{z}_i, \hat{b}))$. A similar argument holds for the other conditions in (e). We next prove that condition (d) holds. First, note that

$$\begin{aligned} \gamma_i = \left(\tfrac{\partial L_i(\hat{y}_i)}{\partial \hat{y}_i} - \tfrac{\partial L_i(\hat{z}_i)}{\partial \hat{y}_i} \right)_+ &= \left(\tfrac{\partial L_i(\max(\hat{a}, \min(\hat{z}_i, \hat{b})))}{\partial \hat{y}_i} - \tfrac{\partial L_i(\hat{z}_i)}{\partial \hat{y}_i} \right)_+ \\ &= \left(\tfrac{\partial L_i(\hat{a})}{\partial \hat{y}_i} - \tfrac{\partial L_i(\hat{z}_i)}{\partial \hat{y}_i} \right)_+ \end{aligned}$$

by convexity of $L_i(\cdot)$ and since $\hat{z}_i^* \geq \hat{a} \Rightarrow \hat{z}_i^* \geq \max(\hat{a}, \min(\hat{z}_i, \hat{b}))$ and $\hat{z}_i^* < \hat{a} \Rightarrow \hat{a} = \max(\hat{a}, \min(\hat{z}_i, \hat{b})) \geq \hat{z}_i^*$. A similar argument shows that $\delta_i = (\partial L_i(\hat{z}_i^*)/\partial \hat{y}_i - \partial L_i(\hat{b})/\partial \hat{y}_i)_+$. Then conditions (d) are satisfied by choosing $\hat{a}$ and $\hat{b}$ to solve equations (14). Conditions (a),(d),(f), and (g) hold by the above constructions. Conditions (b) and (c) hold by construction of $\hat{y}$ in (13). The optimality conditions are satisfied implying the theorem holds true. □

Acknowledgements

The authors thank Rob Tibshirani for a motivating discussion, as well as an anonymous reviewer and associate editor for their useful comments. SR was partially supported by Israeli Science Foundation grant 1487/12.

References

[1] Bacchetti, P. (1989). Additive isotonic model. *Journal of the American Statistical Association 84*(405), 289–294. MR0999691

[2] Barlow, R. and H. Brunk (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association 67*(337), 140–147. MR0314205

[3] Block, H., S. Qian, and A. Sampson (1994). Structure algorithms for partially ordered isotonic regression. *Journal of Computational and Graphical Statistcs 3*(3), 285–300. MR1292119

[4] Chakravarti, N. (1989). Bounded isotonic median regression. *Computational Statistics & Data Analysis 8*, 135–142. MR1016241

[5] Chambolle, A. and P.-L. Lions (1997). Image recovery via total variation minimization and related problems. *Numerische Mathematik 76*(2), 167–188. MR1440119

[6] Chen, X., Q. Lin, and B. Sen (2015). On degrees of freedom of projection estimators with applications to multivariate shape restricted regression. *arXiv:1509.01877*.

[7] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association 81*(394), 461–470. MR0845884

[8] Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*(2), 407–499. MR2060166

[9] Fang, Z. and N. Meinshausen (2012). Liso isotone for high-dimensional additive isotonic regression. *Journal of Computational and Graphical Statistics 21*(1), 72–91. MR2913357

[10] Flynn, C. J., C. M. Hurvich, and J. S. Simonoff (2013). Selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association 108*(503), 1031–1043. MR3174682

[11] Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics 19*(1), 1–67. MR1091842

[12] Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning* (2 ed.). Springer. MR2722294

[13] He, X., P. Ng, and S. Portnoy (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society. Series B 60*(3), 537–550. MR1625950

[14] Hochbaum, D. S. and M. Queyranne (2003). Minimizing a convex cost closure set. *SIAM Journal of Discrete Mathematics 16*(2), 192–207. MR1982135

[15] Hu, X. (1999). Application of the limit of truncated isotonic regression in optimization subject to isotonic and bounding constraints. *Journal of Multivariate Analysis 71*, 56–66. MR1721959

[16] Kato, K. (2009). On the degrees of freedom in shrinkage estimation. *Journal of Multivariate Analysis 100*(7), 1338–1352. MR2514133

[17] Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 20*(5), 1356–1378. MR1805787

[18] Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*(1), 1–27. MR0169712

[19] Lee, C.-I. C. (1983). The min-max algorithm and isotonic regression. *The Annals of Statistics 11*(2), 467–477. MR0696059

[20] Luss, R. and S. Rosset (2014). Generalized isotonic regression. *Journal of Computational and Graphical Statistics 23*(1), 192–210. MR3173767

[21] Luss, R., S. Rosset, and M. Shahar (2012). Efficient regularized isotonic regression with application to gene-gene interaction search. *Annals of Applied Statistics 6*(1), 253–283. MR2951537

[22] Mammen, E. and S. van der Geer (1997). Locally adaptive regression splines. *Annals of Statistics 25*(1), 387–413. MR1429931

[23] Maxwell, W. and J. Muckstadt (1985). Establishing consistent and realistic reorder intervals in production-distribution systems. *Operations Research 33*(6), 1316–1341.

[24] Meyer, M. and M. Woodroofe (2000). On the degrees of freedom in shape-restricted regression. *Annals of Statistics 28*(4), 1083–1104. MR1810920

[25] Obozinski, G., G. Lanckriet, C. Grant, M. Jordan, and W. Noble (2008). Consistent probabilistic outputs for protein function prediction. *Genome Biology 9*, 247–254. Open Access.

[26] Osborne, M. R., B. Presnell, and B. A. Turlach (1999). On the lasso and its dual. *Journal of Computational and Graphical Statistics 9*, 319–337. MR1822089

[27] Rosset, S., G. Swirszcz, N. Srebro, and J. Zhu (2007). L1 regularization in infinite dimensional feature spaces. *Proceedings of the Conference on Learning Theory (COLT)*.

[28] Roundy, R. (1986). A 98%-effective lot-sizing rule for a multi-product, multi-stage productoin/inventory system. *Mathematics of Operations Research 11*(4), 699–727. MR0865565

[29] Schell, M. and B. Singh (1997). The reduced monotonic regression method. *Journal of the American Statistical Association 92*(437), 128–135.

[30] Spouge, M., H. Wan, and W. J. Wilbur (2003). Least squares isotonic regression in two dimensions. *Journal of Optimization Theory and Applications 117*(3), 585–605. MR1989929

[31] Tibshirani, R., H. Hoefling, and R. Tibshirani (2011). Nearly-isotonic regression. *Technometrics 53*(1), 54–61. MR2791946

[32] Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563. MR2274449

[33] Zou, H., T. Hastie, and R. Tibshirani (2007). On the degrees of freedom of the lasso. *Annals of Statistics 35*(3), 2173–2192. MR2363967