Isotonic Recursive Partitioning

Ronny Luss* Saharon Rosset† Moni Shahar ‡

March 1, 2011

Abstract

Isotonic regression is a nonparametric approach for fitting monotonic models to data that has been widely studied from both theoretical and practical perspectives. However, this approach encounters computational and statistical overfitting issues in higher dimensions. To address both concerns we present an algorithm, which we term Isotonic Recursive Partitioning (IRP), for isotonic regression based on recursively partitioning the covariate space through solution of progressively smaller "best cut" subproblems. This creates a regularized sequence of isotonic models of increasing model complexity that converges to the global isotonic regression solution. The models along the sequence are often more accurate than the unregularized isotonic regression model because of the complexity control they offer. We quantify this complexity control through estimation of degrees of freedom along the path. Success of the regularized models in prediction and IRP's favorable computational properties are demonstrated through a series of simulated and real data experiments. We discuss application of IRP to the genetic problem of modeling gene interactions and epistasis, where it appears especially promising.

1 Introduction

In predictive modeling we are given a set of n data observations $(x_1, y_1), ..., (x_n, y_n)$, where $x \in \mathcal{X}$ (usually $\mathcal{X} = \mathbb{R}^d$) is a vector of covariates or independent variables, $y \in \mathbb{R}$ is the response, and we wish to fit a model $\hat{f} : \mathcal{X} \to \mathbb{R}$ to describe the dependence of y on x, i.e., $y \approx \hat{f}(x)$. Isotonic regression is a non-parametric modeling approach which only restricts the fitted model to being monotone in all independent variables (Barlow & Brunk 1972). Define $\mathcal{G}$ to be the family of isotonic functions, that is, $g \in \mathcal{G}$ satisfies

$$x_1 \preceq x_2 \Rightarrow g(x_1) \leq g(x_2),$$

where the partial order $\preceq$ here will usually be the standard Euclidean one, i.e., $x_1 \preceq x_2$ if and only if $x_{1j} \leq x_{2j}$ coordinate-wise. Given these definitions, isotonic regression solves

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

We denote by $\hat{f}$ the optimal solution to (1). As many authors have noted, $\hat{f}$ comprises a partitioning of the space $\mathcal{X}$ into regions with no "holes" satisfying isotonicity properties defined below, with a constant fitted to $\hat{f}$ in every region.

*Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel. ronnyluss@gmail.com
†Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel. saharon@post.tau.ac.il
‡Department of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel. moni@eng.tau.ac.il

In terms of model form, isotonic regression is clearly very attractive in situations where monotonicity is a reasonable assumption, but other common assumptions like linearity or additivity are not. Indeed, this formulation has found useful applications in biology (Obozinski et al. 2008), medicine (Schell & Singh 1997), statistics (Barlow & Brunk 1972) and psychology (Kruskal 1964), among others. In recent years, an exciting new application area has emerged for this approach in genetics: modeling genetic heritability. Many papers have noted the apparent insufficiency of standard additive modeling approaches in describing the combined effects of genetic factors (e.g., mutations) on phenotypes or traits like height (Goldstein 2009, Eichler et al. 2010). In some cases, evidence has pointed to sub-additive interactions (Shao et al. 2008), while others suggest requiring super-additive assumptions in order to explain heritability (Goldstein 2009). It is generally accepted, however, that while the effect of one genetic factor on a phenotype can be modulated, enhanced or even eliminated by other genetic factors, it is not expected to reverse direction (Mani et al. 2007, Roth et al. 2009). In other words, the isotonicity assumption with respect to genetic effects is widely accepted, but the form of epistasis (genetic interaction) between factors is not clear and may vary between phenotypes. Other properties of this application domain also favor the use of isotonic regression as we discuss below.

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, m, and d will refer to these quantities throughout the paper.

The first concern is statistical difficulty and 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 (1989) and Schell & Singh (1997)). At the extreme, there may be no isotonicity constraints because no two observations obey the coordinate-wise requirement for the $\preceq$ ordering. The isotonic solution in this case simply assigns $\hat{f}(x_i) = y_i$ providing a perfect interpolation of the training data. As demonstrated in the literature (Schell & Singh 1997) 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 $\mathcal{G}$, could offer an approach that maintains isotonicity while controlling variance, leading to improved accuracy.

A second concern is computational difficulty. The discussion of isotonic regression originally focused on the case $x \in \mathbb{R}$, where $\preceq$ denoted a complete order (Kruskal 1964). For this case, the well-known pooled adjacent violators algorithm (PAVA) efficiently solves (1) in computational complexity $O(n)$. Low complexities can also be found when the isotonic constraints take a special structure such as a tree ($O(n \log n)$ in Pardalos & Xue (1999)). Various algorithms have been developed for the partially ordered case, including the classical approach of Dykstra & Robertson (1982) for data on a grid, generalizations of PAVA (Lee 1983, Block et al. 1994) and active set methods (de Leeuw et al. 2009). These approaches offer no polynomial complexity guarantees and by all accounts are impractical when data sizes exceed a few thousand observations (in some cases much less). Interior point methods offer complexity guarantees of $O(\max(m,n)^3)$ (Monteiro & Adler 1989), however they are impractical for large data sizes due to excessive memory requirements.

A much more computationally attractive approach can be found in the optimization and operations research literature. The basic idea of this approach is to repeatedly and "optimally" split the covariate space $\mathcal{X}$ into regions of decreasing size by solving a sequence of specially structured *best cut* problems for which efficient algorithms exist. At most n partitions are needed, leading to a computational complexity bounded by $O(n^4)$, and in some cases even less. From a practical performance perspective, this algorithm

can obtain an exact solution of (1) for datasets with tens of thousands of observations in minutes. The first appearance of this approach, to our knowledge, is in the work of Maxwell & Muckstadt (1985) (and similarly Roundy (1986)), who designed it for a problem with a different loss function than in (1) that also had isotonicity constraints. Applicability of their methods with minimal changes to problem (1) was more recently noticed by several authors (e.g. Spouge et al. (2003)), who used it to design efficient isotonic regression algorithms. This approach does not appear to be well-known in the statistics community, and indeed we have independently developed it before discovering it is already known.

The literature cited above invariably refers to this iterative splitting algorithm merely as an approach for efficiently arriving at the optimal solution of (1). However, as noted before, this solution can be highly overfitted, especially as the dimension d increases. Our main interest lies in analyzing the iterative approach as a means towards resolving the overfitting problem, as well as the computational issue. We propose to view this iterative algorithm as a *recursive partitioning* approach that generates isotonic models of increasing model complexity, ultimately leading to the solution of (1); the algorithm is termed Isotonic Recursive Partitioning (IRP). We prove that the models generated by the IRP iterations are indeed isotonic (Theorem 4) and consider them as a *regularization path* of increasingly complex isotonic regression models. Models along the path are less complex, and hence likely to be less overfit and offer better predictive performance than the overall solution to (1), while still maintaining isotonicity. This is confirmed by our analysis of the equivalent degrees of freedom along the IRP path, as well as experiments with simulated and real data.

We observe that for very low dimension (typically $d \leq 2$) the non-regularized solution of (1) performs well. As the dimension increases, regularization becomes necessary, and intermediate models on the IRP path perform better than the non-regularized solution. However, eventually overfitting plagues IRP from its first iteration, and the isotonic models fail to perform better than simple linear regression in out-of-sample prediction, even when the linear model is inappropriate. In our simulations this occurs around dimensions 6-8 even for relatively large data sets.

Progress of IRP is illustrated in Figure 1, where we show an example of applying IRP to the well-known Baseball dataset (He et al. 1998) describing the dependence of salary on a collection of player properties. We limit the model to only two covariates to facilitate visualization, and we choose to use the number of runs batted in and hits since they seemed a-priori most likely to comply with the isotonicity assumptions. The increasing model complexity can be seen, moving from iteration 1 (a single split) through 10 iterations of IRP, to the final isotonic model optimally solving (1), comprising a splitting of the covariate space into 29 regions, each of which is fitted with a constant.

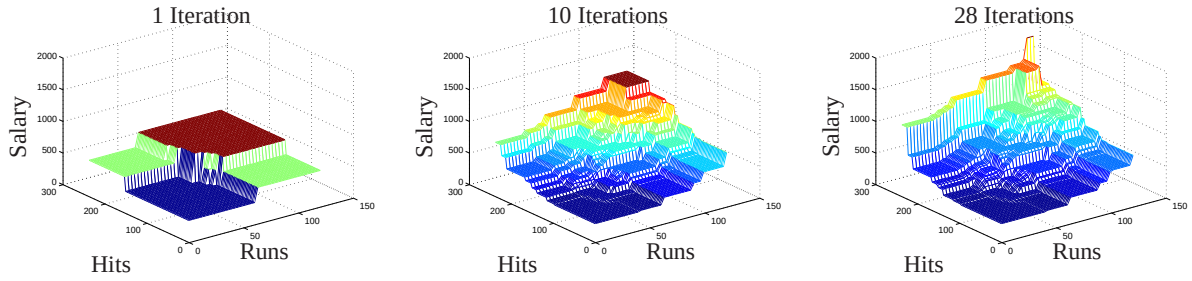

Figure 1: Illustration of IRP on Baseball data. Salary is modeled by number of runs batted in and hits. Models after iterations 1 and 10 of IRP and the final model are shown.

An obvious analogy of IRP can be made to well-known recursive partitioning approaches for regression

such as CART (Breiman et al. 1984), where the iterative splitting of the covariate space generates a sequence of models (trees) of increasing model complexity, from which the "best" tree is chosen via cross validation (for example, using the 1-SE rule (Breiman et al. 1984)). As with CART and other similar approaches, IRP performs a greedy search and finds a "local" optimum in every iteration. However, unlike CART, which has no guarantees on the overall model it generates, IRP is proven to terminate in the global solution of the isotonic regression problem (1). Another difference is that IRP splits are not made along one axis at a time, but rather each split is a non-parametric division of one region in $\mathcal{X}$ into two sub-regions.

The remainder of this paper is organized as follows. We first present and analyze the IRP algorithm in Section 2. We detail the best cut problem solved for splitting at each iteration, and prove that this algorithm is a *no-regret* algorithm, in the sense that it only partitions the data and never merges back previously made partitions and converges to the global solution of (1) (Theorem 2). Furthermore, we prove that the intermediate partitions generated along the IRP path are also isotonic, in the sense that fitting the average to each region gives a model that is in the class $\mathcal{G}$ of isotonic functions in $\mathbb{R}^d$ (Theorem 4). Section 3 briefly reviews the theoretical computational guarantees of IRP as reflected in the literature, and develops a simple and realistic case where the overall computation is $O(n^3)$. Section 4 discusses the statistical model complexity of models generated along the regularization path. Meyer & Woodroofe (2000) have shown that the number of partitions in the solution of (1) is an unbiased estimator of the (equivalent) degrees of freedom (as defined by Efron (1986)). Since IRP adds one to the number of partitions at each iteration, the number of iterations may be used as a parametrization of this sequence. However, we argue that the number of regions is not a good estimate of degrees of freedom because IRP performs much more fitting in its initial iterations compared to later stages, and demonstrate this effect empirically through simulation. We also show that when the covariates are ternary (as is natural in our motivating genetic example when dealing with ternary genotype data), the overall number of degrees of freedom and model complexity increase more slowly with dimension, compared to general continuous covariates, resulting in much less overall fitting for each dimension. Section 5 examines IRP's statistical and computational performance on simulated and real data, specifically pointing out the effect of regularization and increased dimensionality on predictive performance. We apply IRP to simulations with ternary covariates and sub- and super-additive interactions motivated by the genetic application and demonstrate its favorable performance. Section 6 concludes with extensions and connections to previous literature.

We next define terminology to be used throughout the paper.

1.1 Definitions

Let $V = \{x_1, \ldots, x_n\}$ be the covariate vectors for n training points where $x_i \in \mathbb{R}^d$ and denote $y_i \in \mathbb{R}$ as the i^{th} observed response. We will refer to a general subset of points $A \subseteq V$ with no holes (i.e. $x \preceq y \preceq z$ and $x, z \in A \Rightarrow y \in A$) as a *group*. Throughout the paper, we will use the shorthand $x \in A = \{i : x_i \in A\}$. Denote by $|A|$ the cardinality of group A. The *weight* of group A is defined as $\overline{y}_A = \frac{1}{|A|}\sum_{i \in A} y_i$. For two groups A and B, we denote $A \preceq B$ if $\exists x \in A, y \in B$ such that $x \preceq y$ and $\nexists x \in A, y \in B$ such that $y \prec x$ (i.e. there is at least one comparable pair of points that satisfy the direction of isotonicity). A set of groups $\mathcal{V}$ is called isotonic if $A \preceq B \Rightarrow \overline{y}_A \leq \overline{y}_B, \forall A, B \in \mathcal{V}$. The groups within this set $\mathcal{V}$ are referred to as isotonic regions. A subset $\mathcal{L}$ ($\mathcal{U}$) of A is a *lower set* (*upper set*) of A if $x \in A, y \in \mathcal{L}, x \prec y \Rightarrow x \in \mathcal{L}$ ($x \in \mathcal{U}, y \in A, x \prec y \Rightarrow y \in \mathcal{U}$).

A group $B \subseteq A$ is defined as a block of group A if $\overline{y}_{\mathcal{U} \cap B} \leq \overline{y}_B$ for each upper set $\mathcal{U}$ of A such that $\mathcal{U} \cap B \neq \{\}$ (or equivalently if $\overline{y}_{\mathcal{L} \cap B} \geq \overline{y}_B$ for each lower set $\mathcal{L}$ of A such that $\mathcal{L} \cap B \neq \{\}$). A set of blocks $\mathcal{S} = \{B_1, \ldots, B_k\}$ is called *block class* of V if $B_i \cap B_j = \{\}$ and $B_1 \cup \ldots \cup B_k = V$. $\mathcal{S}$ is an *isotonic block class* if $\forall B_i, B_j \in \mathcal{S}$, $B_i \preceq B_j \Rightarrow \overline{y}_{B_i} \leq \overline{y}_{B_j}$. A group X *majorizes* (*minorizes*) another group Y if

$X \succeq Y$ ($X \preceq Y$). A group X is a *majorant* (*minorant*) of $X \cup A$ where $A = \cup_{i=1}^{k} A_i$ if $X \not\prec A_i$ ($X \not\succ A_i$) $\forall i = 1 \dots k$.

We denote the optimal solution for minimizing $f(x)$ in the variable x by x^*, i.e. $x^* = \operatorname{argmin} f(x)$.

2 IRP and a regularization path for isotonic regression

We describe here the partitioning algorithm used to solve the isotonic regression problem (1). Section 2.1 first reformulates the isotonic regression problem and describes the structure of the optimal solution. Section 2.2 motivates and details the IRP algorithm and, in particular, the main partitioning step. Each group created by the partitioning scheme is proven to be the union of blocks in the optimal solution, i.e. all partitions have the no-regret property. An important aspect of the algorithm is the regularization path generated as a byproduct as each partition creates a new feasible solution. Section 2.3 goes on to prove convergence of IRP to the global optimal solution of (1), and most importantly, that each solution along the regularization path is isotonic.

2.1 Structure of the isotonic solution

Isotonic regression seeks a monotonic function that fits a given training dataset $\{x_i, y_i\}_{i=1}^{n}$ and satisfies a set of *isotonicity constraints* which we index by the set $\mathcal{I} = \{(i,j) : x_i \preceq x_j\}$. We will usually assume that $x_i \in \mathbb{R}^d$ and that $\preceq$ is the standard partial order in $\mathbb{R}^d$ based on coordinate-wise inequalities. A reformulation of (1) is

$$\min \{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 : \hat{y}_i \leq \hat{y}_j \quad \forall (i,j) \in \mathcal{I}\}. \tag{2}$$

Problem (2) is a quadratic program with linear constraints. Any solution satisfying the constraints given by $\mathcal{I}$ is referred to as an isotonic, or feasible, solution. The structure of the optimal solution to (2) is well-known: Observations are divided into k groups where the fits in each group take the group mean observation value. This can be seen through the following Karush-Kuhn-Tucker (KKT), i.e. optimality, conditions (Boyd & Vandenberghe 2004) to (2):

(a) $\hat{y}_i = y_i - \frac{1}{2}(\sum_{j:(i,j)\in\mathcal{I}} \lambda_{ij} - \sum_{j:(j,i)\in\mathcal{I}} \lambda_{ji})$

(b) $\hat{y}_i \leq \hat{y}_j \; \forall (i,j) \in \mathcal{I}$

(c) $\lambda_{ij} \geq 0 \; \forall (i,j) \in \mathcal{I}$

(d) $\lambda_{ij}(\hat{y}_i - \hat{y}_j) = 0 \; \forall (i,j) \in \mathcal{I}$,

where λ_{ij} is the dual variable corresponding to the isotonicity constraint $\hat{y}_i \leq \hat{y}_j$. This set of conditions exposes the nature of the optimal solution. Condition (d) implies that $\lambda_{ij} > 0 \Rightarrow \hat{y}_i = \hat{y}_j$ meaning λ_{ij} can be non-zero only within blocks in the isotonic solution which have the same fitted value. For observations in different blocks, $\lambda_{ij} = 0$. Furthermore, the fit within each block is trivially seen to be the average of the observations in the block, as the average minimizes the block's squared loss. A block is thus also referred to as an *optimal group* with respect to an isotonic regression problem. Condition (b) implies isotonicity of the blocks, and thus, we get the familiar characterization of the isotonic regression problem as one of finding a division into an isotonic block class.

2.2 The partitioning algorithm

Suppose a current group V is optimal (i.e. V is a block) and thus the optimal fits at points in V, denoted $\hat{y}_i^*$, satisfy $\hat{y}_i^* = \overline{y}_V$ for all $i \in V$, which leads to the condition $\sum_{i \in V}(y_i - \overline{y}_V) = 0$. Then finding two groups A and B within V such that $\sum_{i \in B}(y_i - \overline{y}_V) - \sum_{i \in A}(y_i - \overline{y}_A) > 0$ should be infeasible, according to the KKT conditions. The division in IRP looks for two such groups. Denote by $\mathcal{C}_V = \{(A, B) : A, B \subseteq V, A \cup B = V, A \cap B = \{\}, \nexists x \in A, y \in B \quad \text{s.t.} \quad y \preceq x\}$ the set of all feasible (i.e. isotonic) partitions defined by observations in V. We refer to partitioning as making a cut through the variable space (hence our optimal partition is made by an *optimal cut*). The optimal cut is determined by the partition that solves the problem

$$\max_{(A,B) \in \mathcal{C}_V} \{\sum_{i \in B}(y_i - \overline{y}_V) - \sum_{i \in A}(y_i - \overline{y}_V)\} = \{-|A|(\overline{y}_A - \overline{y}_V) + |B|(\overline{y}_B - \overline{y}_V)\} \tag{3}$$

where $A(B)$ is the group on the lower (upper) side of the edges of cut. A more statistically intuitive rule might look for the split that maximizes between-group variance. This partitioning problem solves

$$\max_{\{(A,B) \in \mathcal{C}_V\}} \{|A|(\overline{y}_A - \overline{y}_V)^2 + |B|(\overline{y}_B - \overline{y}_V)^2\}. \tag{4}$$

The next proposition makes a connection between the above two maximization problems, and draws a clear conclusion on the relationship between their optimal solutions, namely that the optimal partitions to (3) are always more balanced than the optimal partitions to (4). The next proposition makes a connection between the two maximization problems, and draws a clear conclusion on the relationship between the optimal solutions, namely that the split generated by (3) is always more balanced than the split generated by (4).

Proposition 1 *Denote the optimal solutions of the optimal cut problem (3) and the between-group variance maximization problem (4) by* (A^*, B^*) *and* $(\tilde{A}, \tilde{B})$ *and their objective functions by* $g^*(A, B)$ *and* $\tilde{g}(A, B)$*, respectively. Then*

$$(A^*, B^*) = \underset{(A,B) \in \mathcal{C}_V}{\operatorname{argmax}} \{|A||B|\tilde{g}(A, B)\}$$

and

$$(|A^*| - |B^*|)^2 \leq (|\tilde{A}| - |\tilde{B}|)^2.$$

We leave the proof to the appendix.

Thus, we can look at the IRP criterion as a modified form of maximizing between-group variance which encourages more balanced splitting. However, while solving the partition problem (4) is difficult, the IRP partition problem (3) is tractable. Indeed, the optimal partition problem (3) can be reduced to solving the linear program

$$\max \{z^T x : x_i \leq x_j \quad \forall (i, j) \in \mathcal{I}, -1 \leq x_i \leq 1 \quad \forall i \in V\} \tag{5}$$

where $z_i = y_i - \overline{y}_V$. If the optimal objective value equals zero, then the group V must be an optimal block.

This group-wise partitioning operation is the basis for our IRP algorithm which is detailed in Algorithm 1. It starts with all observations as one group and recursively splits each group optimally by solving sub-problem (5). At each iteration, a list $\mathcal{C}$ of potential optimal partitions for each group generated thus far is maintained, and the partition among them with the highest objective value is performed. The list $\mathcal{C}$ is updated with the optimal partitions generated from both sub-groups. Partitioning ends whenever the solution to (5) is trivial (i.e., no split is found because the group is a block). We can think of each iteration k of Algorithm

1 as producing a model M_k by fitting the average to each group in its current partition: For a set of groups $\mathcal{V} = \{V_1, \ldots, V_k\}$, denote $\overline{y}_{\mathcal{V}} = \{\overline{y}_{V_1}, \ldots, \overline{y}_{V_k}\}$. Then model $M_k = (\mathcal{V}, \overline{y}_{\mathcal{V}})$ contains the partitioning $\mathcal{V}$ as well as a fit to each of the observations, which is the mean observation of the group it belongs to in the partition.

Algorithm 1 Isotonic Recursive Partitioning

Require: Observations $y_1, \ldots, y_n$ and partial order $\mathcal{I}$.
Require: $\mathcal{A} = \{\{1, \ldots, n\}\}$,$\mathcal{C} = \{(0, \{1, \ldots, n\}, \{\})\}$,$\mathcal{B} = \{\}$, $M_0 = (\mathcal{A}, \overline{y}_{\mathcal{A}})$.
while $A \neq \{\}$ **do**
  Let $(val, w^-, w^+) \in \mathcal{C}$ be the potential partition with largest val.
  Update $\mathcal{A} = (\mathcal{A} \setminus (w^- \cup w^+)) \cup \{w^-, w^+\}$, $\mathcal{C} = \mathcal{C} \setminus (val, w^-, w^+)$.
  $M_k = (\mathcal{A} \cup \mathcal{W}, \overline{y}_{\mathcal{A} \cup \mathcal{B}})$.
  for all $v \in \{w^-, w^+\}$ **do**
    Set $z_i = y_i - \overline{y}_v \; \forall i \in v$ where $\overline{y}_v$ is the mean of observations indexed by v.
    Solve LP (5) with input z and get $x^* = \operatorname{argmin} \mathrm{LP}(5)$.
    if $x_1^* = \ldots = x_n^*$ (group is optimally divided) **then**
      Update $\mathcal{A} = \mathcal{A} \setminus v$ and $\mathcal{B} = \mathcal{B} \cup v$.
    else
      Let $v^- = \{i : x_i^* = -1\}, v^+ = \{i : x_i^* = +1\}$.
      Update $\mathcal{C} = \mathcal{C} \cup \{(z^T x^*, v^-, v^+)\}$
    end if
  end for
end while
return $\mathcal{B}$, indices of observations corresponding to the optimal groups.

Note: Sets here keep track of indices rather than observations for ease of implementation.

2.3 Properties of the partitioning algorithm

Theorem 2 next states the result which implies that the IRP partitions are *no-regret*. This will lead to our convergence result.

Theorem 2 *Assume group V is the union of blocks from the optimal solution to problem (2). Then a cut made by solving (3) (using 5) at a particular iteration does not cut through any block in the global optimal solution.*

The fact that IRP is a no-regret algorithm can be shown using a connection between the work of Barlow & Brunk (1972) and Maxwell & Muckstadt (1985) (held to the Discussion in Section 6). We prove Theorem 2 directly, but leave it to the Appendix as the theorem is already known to be true (Spouge et al. 2003). Remark 7 in the Appendix handles the case for multiple observations. Since Algorithm 1 starts with $\mathcal{A} = \{1, ..., n\}$ which is the union of all blocks, we can conclude from this theorem that IRP never cuts an optimal block when generating partitions. The following corollary is then a direct consequence of repeatedly applying Theorem 2 in Algorithm 1:

Corollary 3 *Algorithm 1 converges to the optimal (isotonic block class) solution with no regret.*

Theorem 4 next states our main innovative result that Algorithm 1 provides isotonic solutions at each iteration. This result implies that the path of solutions generated by IRP can be regarded as a regularization path for isotonic regression. Along the path, the model grows in complexity until optimality. These suboptimal isotonic models often result in better predictive performance than the optimal solution, which is susceptible to overfitting as is discussed in Section 5.

Theorem 4 *Model M_k generated after iteration k of Algorithm 1 is in the class $\mathcal{G}$ of isotonic models.*

Proof.

The proof is by induction. The base case, i.e. first iteration, where all points form one group is trivial. The first cut is made by solving the linear program (5) which constrains the solution to maintain isotonicity.

Assuming that iteration k (and all previous iterations) provides an isotonic solution, we prove that iteration $k+1$ must also maintain isotonicity. Figure 2 helps illustrate the situation described here. Let G be the group split at iteration $k+1$ and denote A (B) as the group under (over) the cut. Let $\mathcal{A} = \{X : X$ is a group at iteration $k+1, \exists i \in X$ such that $(i,j) \in \mathcal{I}$ for some $j \in A\}$ (i.e. $X \in \mathcal{A}$ border A from below).

Consider iteration $k+1$. Denote $\mathcal{X} = \{X \in \mathcal{A} : \overline{y}_A < \overline{y}_X\}$ (i.e. $X \in \mathcal{X}$ violates isotonicity with A). The split in G causes the fit in nodes in A to decrease. We will prove that when the fits in A decrease, there can be no groups below A that become violated by the new fits to A, i.e. the decreased fits in A cannot be such that $\mathcal{X} \neq \{\}$.

We first prove that $\mathcal{X} = \{\}$ by contradiction. Assume $\mathcal{X} \neq \{\}$. Denote $i < k+1$ as the iteration at which the last of the groups in $\mathcal{X}$, denoted D, was split from G and suppose at iteration i, G was part of a larger group H and D was part of a larger group F. It is important to note that $X \bigcap (F \bigcup H) = \{\}$ $\forall X \in \mathcal{X} \setminus D$ at iteration i because by assumption all groups in $\mathcal{X} \setminus D$ were separated from A before iteration i. Thus, at iteration i, D is the only group bordering A that violates isotonicity.

Let D_U denote the union of D and all groups in F that majorize D. By construction, D_U is a majorant in F. Hence $\overline{y}_{D_U} < \overline{y}_{F \cup H}$ by Algorithm 1 and $\overline{y}_A < \overline{y}_{D_U}$ by definition since $\overline{y}_{D_U} > \overline{y}_D > \overline{y}_A$. Also by construction, any set $X \in H$ that minorizes A has $\overline{y}_X < \overline{y}_A$ (each set X that minorizes A besides D such that $\overline{y}_X < \overline{y}_A$ has already been split from A). Hence we can denote A_L as the union of A and all groups in H that minorize A and we have $\overline{y}_A > \overline{y}_{A_L}$ and A_L is a minorant in H. Since $A_L \subseteq H$ at iteration i, we have

$$\overline{y}_{F \cup H} < \overline{y}_{A_L} < \overline{y}_A < \overline{y}_{D_U} < \overline{y}_{F \cup H}$$

which is a contradiction, and hence the assumption $\mathcal{X} \neq \{\}$ is false. The first inequality is because the algorithm left A_L in H when F was split from H, and the remaining inequalities are due to the above discussion. Hence the split at iterations $k+1$ could not have caused a break in isotonicity.

A similar argument can be made to show that the increased fit for nodes in B does not cause any isotonic violation. The proof is hence completed by induction. ■

With Theorem 4, the machinery for generating a regularization path is complete. In Section 3, we describe the computational complexity for generating this path followed by a discussion of the statistical complexity of the solutions along the path in Section 4.

3 Complexity

We here show that the partitioning step in IRP can be solved efficiently. The computational bottleneck of Algorithm 1 is solving linear program (5) that iteratively partitions each group. Linear program (5) has

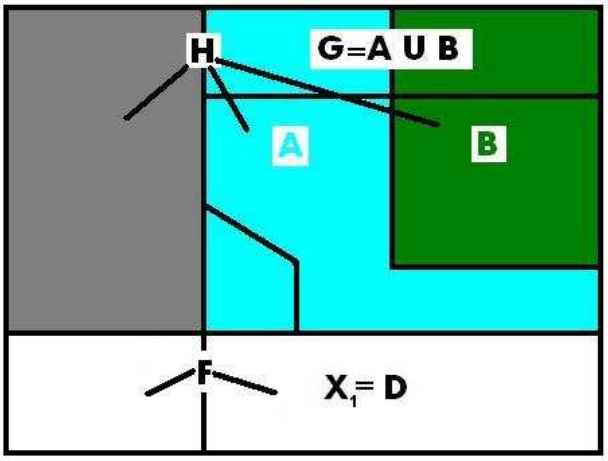

Figure 2: Illustration of proof of Theorem 4 showing the defined sets at iteration $k+1$. G is the set divided at iteration $k+1$ into A (all blue area) and B (all green area). The group bordering A from below denoted by X_1 (also referred to as D in the proof) is in violation with A. At iteration k_0, G is part of the larger group H and X_1 is part of the larger group F. At iteration k_0, groups F and H are separated. The proof shows that when A and B are split at iteration $k+1$, no group such as X_1 where $w_{X_1} > w_A$ could have existed. In the picture, X_1 must have been separated at an iteration $k_0 < k+1$, but the proof, through contradiction, shows that this cannot occur.

a special structure that can be taken advantage of in order to solve larger problems faster. Indeed, the dual problem can be written as an optimization problem called a network flow problem that is amenable to very efficient algorithms, as noted by Spouge et al. (2003) who recognize the network flow problem as the *maximal upper set problem*. We note that our partition problem (5) is very similar to the network flow problem solved in Chandrasekaran et al. (2005) where z_i there represents the classification performance on node i.

We denote the complexity of solving linear program (5) by $T(m,n)$ where m is the number of constraints defined by $\mathcal{I}$ and n is the number of observations. Various efficient algorithms for solving this problem exist, giving complexities such as $T(m,n) = O(mn \log n)$ (Sleator & Tarjan 1983) along with several algorithms giving $T(m,n) = O(n^3)$ (Galil & Naamad 1980). Choosing the more efficient implementation depends on the number of isotonicity constraints m (e.g. $n^3 \leq mn \log n$ for the worst case $m = O(n^2)$). A recent result by Stout (2010) shows how to represent an isotonic regression problem by an equivalent problem where both the number of total observations and constraints are of order $O(n \log^{d-1} n)$, which greatly reduces the worst case of $m = O(n^2)$ isotonicity constraints (i.e. by trading off a few additional *shadow* observations for a large reduction in the number of constraints). Since at most n partitions are made by IRP, complexity is $O(n^4)$ using $T(m,n) = O(n^3)$ or reduced to $O(n^3 \log^{2d-1} n)$ using results of Stout (2010).

In practice, the complexity can be even better by accounting for the fact that IRP solves a sequence of partitioning problems that are decreasing in size (i.e. problems with fewer and fewer observations). Each partition in Algorithm 1 can be divided into different proportions. We generically denote the bigger proportion in a partition by $p \geq 0.5$. Proposition 5 next gives a bound on the complexity of Algorithm 1 for this general case (assuming $T(m,n) = O(n^3)$), in terms of the maximal p over all partitions.

Proposition 5 *Let $p_{\max} \geq .5$ be the greatest p over all iterations of Algorithm 1 such that iteration k partitions a group of size n_k into two groups of size pn_k and $(1-p)n_k$. Denote by n the total number of observations. Then the complexity of Algorithm 1 is bounded by*

$$O(n^3)\frac{1}{1-p_{\max}^2}. \tag{6}$$

The proof, given in the Appendix, is based on the fact that the sequence of IRP's partition problems are solved on smaller and smaller groups of observations (i.e. while the first partition problem is $O(n^3)$, the partition problems for the two created partitions are $O(p^3n^3)$ and $O((1-p)^3n^3)$ for some p where $0 < p < 1$). Even at $p_{\max} = .99$, the constant $1/(1-p_{\max}^2) \approx 50$, which is very small when the number of observations is large. Thus, under another reasonable assumption that $p_{\max}$ is bounded, we can conclude that IRP is of practical complexity $O(n^3)/(1-p_{\max}^2)$. Similar analysis using results of Stout (2010) lead to a practical complexity of $O(n^2 \log^{2d-1} n)/(1-p_{\max})$.

4 Degrees of freedom of isotonic regression and IRP

The concept of degrees of freedom is commonly used in statistics to measure the complexity of a model (or more accurately, a modeling approach). This concept captures the amount of fitting the model performs, as expressed by the optimism of the in-sample error estimates, compared to out-of-sample predictive performance. Here we briefly review the main ideas of this general approach, and then apply them to isotonic regression and IRP.

Following Efron (1986) and Hastie et al. (2001), assume the values $x_1, ..., x_n \in \mathbb{R}^d$ are fixed in advance (the *fixed-x* assumption), and that the model gets one vector of observations $\mathbf{y} = (y_1, ..., y_n)^\intercal \in R^n$ for training, drawn according to $P(y|x)$ at the n data points. Denote by $\mathbf{y}^{\text{new}}$ another independent vector drawn according to the same distribution. $\mathbf{y}$ is used for training a model $\hat{f}(x)$ and generating predictions at the n data points $\hat{y}_i = \hat{f}(x_i)$.

We define the *in-sample* mean squared error:

$$\text{MRSS} = \frac{1}{n}\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$$

and compare it to the expected error the same model incurs on the new, independent copy, denoted in Hastie et al. (2001) by ERR_{in}:

$$\text{ERR}_{\text{in}} = \frac{1}{n}\mathbb{E}_{\mathbf{y}^{\text{new}}}\|\mathbf{y}^{\text{new}} - \hat{\mathbf{y}}\|_2^2.$$

The difference between the two is the *optimism* of the in-sample prediction. As Efron (1986) and others have shown, the expected optimism in MRSS is:

$$E_{\mathbf{y},\mathbf{y}^{\text{new}}}(\text{ERR}_{\text{in}} - \text{MRSS}) = \frac{2}{n}\sum_i \text{cov}(y_i, \hat{y}_i). \tag{7}$$

For linear regression with homoskedastic errors with variance σ^2, it is easy to show that (7) is equal to $\frac{2}{n}d\sigma^2$, where d is the number of regressors, hence the degrees of freedom. This naturally leads to defining

the *equivalent degrees of freedom* of a modeling approach as:

$$df = \sum_i \text{cov}(y_i, \hat{y}_i)/\sigma^2. \tag{8}$$

In non-parametric models, one usually cannot calculate the actual degrees of freedom of a modeling approach, but it is often easier to generate *unbiased estimates* $\hat{df}$ of *df* using Stein's lemma (Stein 1981). Meyer & Woodroofe (2000) demonstrate the applicability of this theory in shape-restricted non-parametric regression. Specifically, their Proposition 2, adapted to our notation, implies that if we assume the homoskedastic case $var(y_i) = \sigma^2$ for all i, then the unbiased estimator $\hat{df}$ for degrees of freedom in isotonic regression is the number of pieces D in the solution $\hat{\mathbf{y}}$ to (2), that is:

$$\mathbb{E}(D) = \sum_i \text{cov}(y_i, \hat{y}_i)/\sigma^2.$$

Considering the IRP algorithm, this puts us in the interesting situation where the number of steps the algorithm takes until it terminates in the global isotonic solution is equal to the degrees of freedom estimator of this global solution (minus one, since we start with one piece). One might thus be inclined to assume that each iteration of Algorithm 1 adds *about* one degree of freedom, i.e. performs approximately the same amount of fitting in every iteration. A similar idea is represented by the degrees of freedom calculation of Schell & Singh (1997) in their reduced monotonic regression algorithm (which starts from the complete isotonic fit and eliminates pieces).

On more careful consideration, however, it is obvious that this idea is incorrect since the first iteration of IRP finds an optimal cut in the very large space of all possible multivariate isotonic cuts. For comparison, a single deep split in a regression tree has been estimated to consume three or more degrees of freedom (Ye 1998), and the space of possible splits in initial IRP iterations is much larger than that of a regression tree since IRP splits are not limited to being axis-oriented. Thus, intuitively, the first iteration is expected to use much more than one degree of freedom (the equivalent of fitting one coefficient to a *fixed, predetermined* regressor). This effect should be exacerbated as the dimension d of x increases since the size of the search space for isotonic cuts increases with it. It also inevitably implies that the latter iterations of the IRP algorithm should perform less (ultimately much less) fitting than the equivalent of one degree of freedom in every iteration, to be consistent with the unbiasedness of $\hat{df} = D$ as an estimator of *df*.

Here we demonstrate empirically that this is indeed the case. We simulate data from a simple additive model

$$x_{ij} \sim \mathcal{U}[1, 2] \text{ i.i.d} \tag{9}$$

$$y_i = \sum_j x_{ij}^2 + \mathcal{N}(0, 10). \tag{10}$$

where x_{ij} is dimension j of the observation i. We can repeatedly generate data using (9) and (10), apply IRP, and empirically estimate *df* as defined by (8) for every iteration of IRP. Figure 3 (left) shows how *df* evolves in this model as the IRP iterations proceed, for increasing dimensions of x. The covariance in (8) was estimated by drawing values $X = (x_1, ..., x_{1000})$ according to the model (9), fixing them, drawing 1000 independent copies of $\mathbf{y}|X$ according to (10), and applying IRP on each one. This whole process was repeated 50 times and the results were averaged. As expected, we see that the number of pieces (hence degrees of freedom) in the final isotonic regression increases with the dimension, as does the rate in which the number of degrees of freedom increases in the initial steps of IRP.

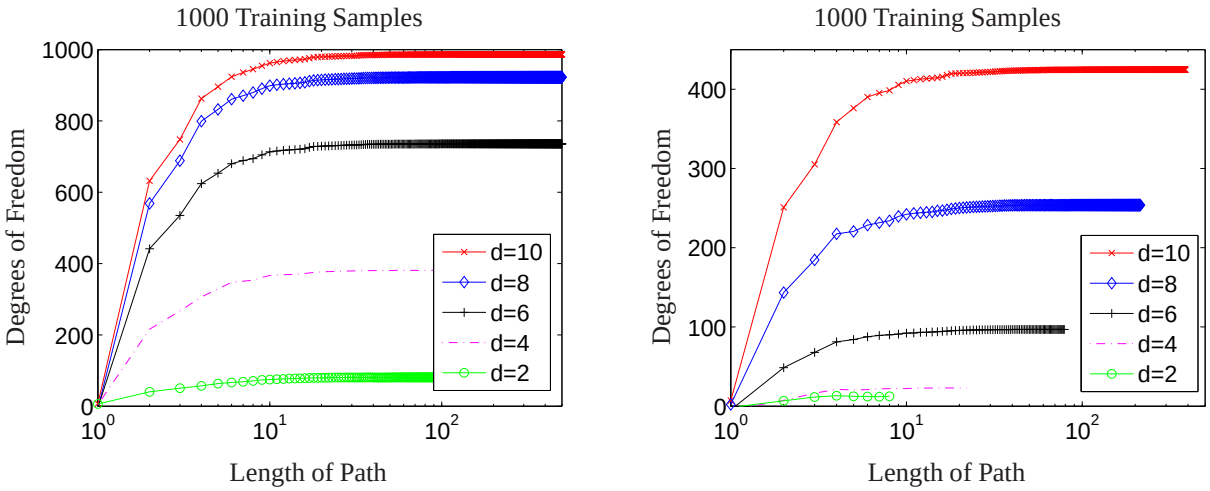

Figure 3: Evolution of degrees of freedom for IRP as model complexity increases. Both models use $y_i = \sum_j x_{ij}^2 + \mathcal{N}(0, 10)$. Simulation (left) uses $x_{ij} \sim \mathcal{U}[1, 2]$ and simulation (right) uses $x_{ij} \in \{0, 1, 2\}$ with probabilities $\{1/3, 1/3, 1/3\}$. Each path is the mean over 50 trials with 1000 training samples.

In order to emphasize this dependence of the degrees of freedom in initial iterations on the dimension, as well as on the number of observations, Figure 4 presents the evolution of the percentage of the total isotonic regression degrees of freedom along the path (i.e., number of degrees of freedom relative to the number of partitions of the final model) as a function of both the dimension and the amount of data used. As expected, increasing the dimension radically increases the portion of the fitting in the first steps, while increasing the amount of data decreases this portion (since the overall isotonic fit is generally more complex in these situations). It should be noted that for many of the situations examined, IRP performs more than half of the total isotonic fit, as measured by degrees of freedom, in its first iteration! In dimension 7, even at $n = 20,000$ observations, almost 60% of the total fitting is associated with the first iteration. Thus, these simulations clearly demonstrate the nature and limitations of IRP's regularization behavior: the IRP path contains models that are regularized isotonic models compared to the global solution, but IRP's ability to control model complexity is limited by the concentration of most of the fitting in the initial iterations, especially in higher dimension.

As mentioned in the introduction, an area that combines applications where the isotonic assumptions are reasonable (i.e., low bias) and the overfitting may be of less concern (i.e., variance can be controlled) is in genetics, specifically in modeling gene-gene interactions in phenotype(y)-genotype(X) relationships (Cordell 2009). The key observation here is that genotypes are ternary ($x_{ij} \in \{0, 1, 2\}$ copies of the "risk" allele). Thus, each dimension of the predictor space $\mathcal{X}$ can take only one of three possible observed values in the data. Intuitively, it is clear that this would significantly reduce the space of possible isotonic splits in IRP, and hence reduce the amount of fitting. To demonstrate this empirically, Figure 3 (right) displays an experiment with the same setup, where instead of drawing the x values from a multivariate uniform, they are drawn independently from $\{0, 1, 2\}$ with equal probabilities and we use the same model. Figure 3 demonstrates that both the globally optimal isotonic regression and, especially, the first IRP iterations perform much less overall fitting in the ternary case versus the continuous case, as measured by equivalent degrees of freedom. For example, in six dimensions, the continuous case requires almost seven times as

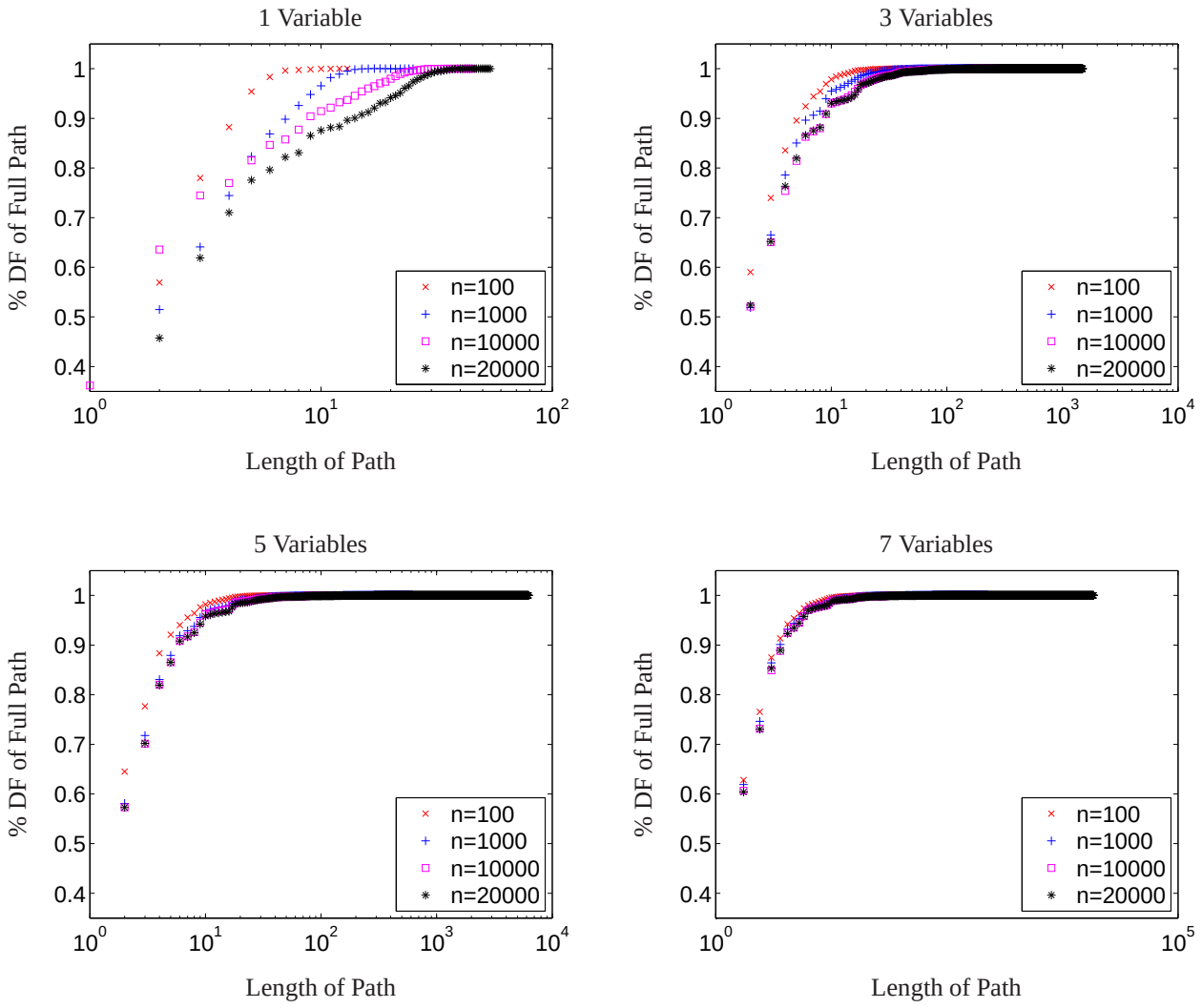

Figure 4: Percentage of degrees of freedom relative to the full path (i.e. number of partitions in the globally optimal solution) as model complexity increases. Simulations use $\mathbf{x}_{ij} \sim \mathcal{U}[1,2]$ with $y_i = \sum_j x_{ij}^2 + \mathcal{N}(0,10)$. Each path is the mean over 200 trials. Only the first 500 partitions of the paths are displayed in order to make the MSE of the earlier IRP iterations visually clearer.

many degrees of freedom than in the ternary case to fit the model. However, relative to the final model, a large percentage of the fitting still takes place in the initial iterations.

5 Performance evaluation

We demonstrate here the usefulness of isotonic regression on simulation and real data. For each experiment, IRP is run on the training data and produces a path of isotonic models. Each model is used for prediction on the test data and the mean squared error (MSE) is recorded. This generates paths of mean squared errors over the different isotonic models and is illustrated in the figures below. In each table, we record the minimum MSE along these paths (*IRP Min MSE*), along with how many partitions were made to generate this mini-

mum MSE (*IRP Min Path*), and the number of partitions in the global isotonic solution (*IRP Path Length*). IRP, as well as optimal isotonic regression, results are compared to running a least squares regression on the training data and predicting on the testing data with the resulting linear model (corresponding MSE is called *LS MSE*), and to the performance of the global isotonic regression solution (*Isotonic Regression MSE*). Because we are interested in examining the behavior of the entire IRP path as in selecting the optimal tuning parameter, and to avoid a significant increase in running time, we do not employ cross validation for selecting the best stopping point (number of iterations), but use test sets for this. In practical application, cross validation would be the appropriate approach for selecting the best model for prediction.

5.1 Simulations

We first illustrate isotonic regression on simulated data with different distributions. For the first three experiments, the i^{th} observation's regressors are distributed as $x_{ij} \sim \mathcal{U}[0,3]$, $x_{ij} \sim \mathcal{U}[0,5]$, and $x_{ij} \sim \mathcal{U}[0,2]$, respectively, and in all cases all x_{ij} are i.i.d. Responses y_i for the three simulations are generated as

$$y_i = (\prod_{j=1}^{d} x_{ij}) + \mathcal{N}(0, d^2), \quad y_i = (\textstyle\sum_j x_{ij}^2) + \mathcal{N}(0, 4d^2), \quad \text{and} \quad y_i = 2^{\sum_j x_{ij}} + \mathcal{N}(0, d^2)$$

respectively, where d is the dimension. The last two experiments are ternary. The i^{th} observation's regressors are distributed as $x_{ij} \in \{0, 1, 2\}$ with probabilities $\{.7, .2, .1\}$ and $\{1/3, 1/3, 1/3\}$ for the fourth and fifth experiments, respectively. The fourth model is subadditive while the fifth model is superadditive; specifically they are

$$y_i = (\textstyle\sum_j x_{ij})^{1/4} + \mathcal{N}(0, d^2/10) \quad \text{and} \quad y_i = (\prod_j x_{ij}) + \mathcal{N}(0, d^2).$$

For each of 50 simulations, 12000 training and 3000 testing points were randomly generated and statistics computed (all tests are out-of-sample).

Figure 5 demonstrates testing error for IRP over the regularized path of isotonic solutions for the first three experiments (with continuous covariates). Each path is normalized by presenting the ratio of model MSE relative to the MSE of a baseline (null) model which fits the mean of the training data. The main observation here is that as the dimension increases, the effect of overfitting of the standard (non-regularized) isotonic regression becomes more significant and causes the skewed u-shaped pattern across the IRP path, where the minimum prediction MSE is obtained earlier in the path. This is the effect we alluded to in the introduction and it stems from the limitations of the isotonicity constraints in controlling model complexity in high dimension.

Table 1 displays certain statistics on all five simulations as well as a comparison to the results of a least squares regression. We first discuss the case of continuous covariates (first three models). In lower dimensions standard isotonic regression performs well, and regularization through IRP offers no gain (this is seen in dimension $d = 2$ in all three examples). Here, isotonic regression controls bias by accommodating the non-linearities in the true model model and significantly outperforms least squares regression. As the number of covariates increases, regularization through IRP becomes necessary to control variance, and the optimal performance is obtained earlier in the IRP path (dimension $d = 4$ in our examples). When d increases further, however, IRP also becomes inefficient at controlling variance, and linear regression dominates. This effect can be traced back to the large amount of fitting performed by IRP already in its initial iterations, as demonstrated in the previous section.

With respect to the models with ternary covariates, isotonic regression outperforms the simple linear regression, however the IRP path does not statistically improve performance. For the subadditive model

(Model 4), performance is better for dimensions 2 and 4, after which again IRP is unsuccessful at controlling variance. However, the superadditive model (Model 5) dominates for all dimensions.

Thus, our simulations confirm that isotonic regression performs well in low dimension, but requires a lot of data in order to learn good nonlinear models in higher dimensions. In intermediate dimension, IRP can offer a compromise between fitting flexible isotonic models and controlling model complexity, resulting in useful prediction models.

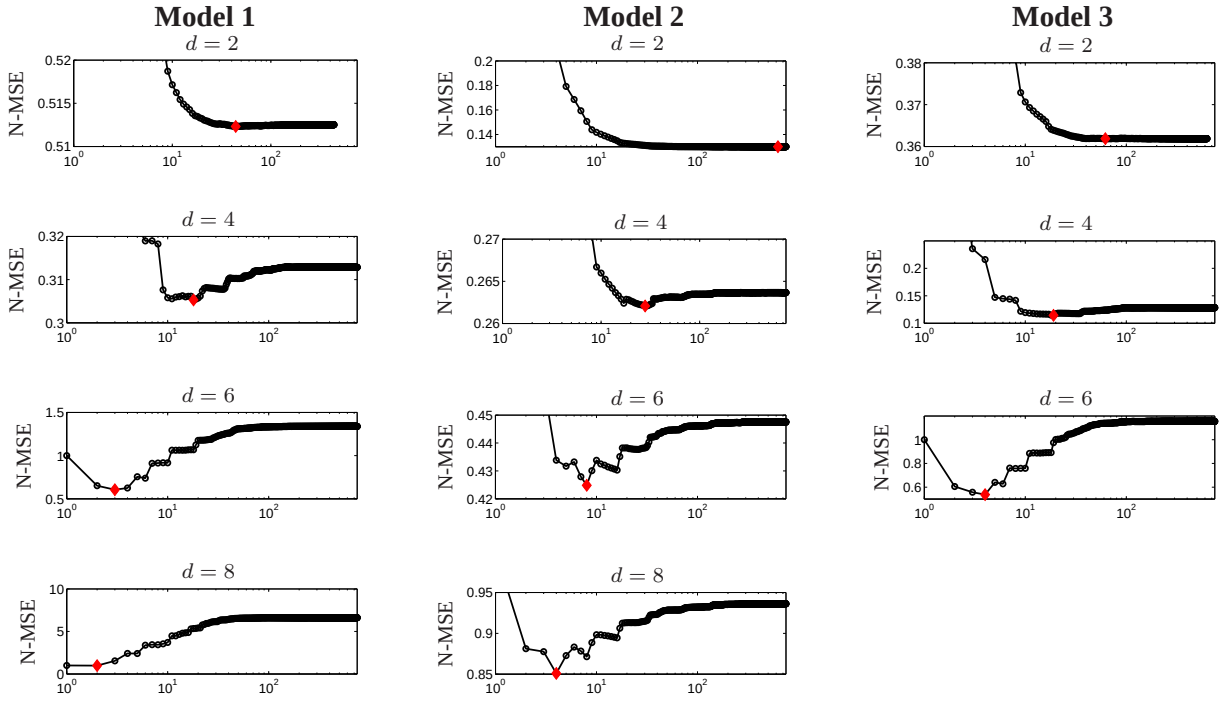

Figure 5: Normalized mean squared error for out-of-sample predictions of simulations with different dimensions d. Each path is normalized by the MSE of the initial model where each training point is fit to the mean. The x-axis in each figure corresponds to the number of partitions made by IRP, i.e. the curves show how the normalized MSE of test data varies as the IRP algorithm progresses. Model 1 uses the function $y_i = (\prod_j x_{ij}) + \mathcal{N}(0, d^2)$ with $x_{ij} \sim \mathcal{U}[0, 3]$. Model 2 uses the function $y_i = (\sum_j x_{ij}^2) + \mathcal{N}(0, 4d^2)$ with $x_{ij} \sim \mathcal{U}[0, 5]$. Model 3 uses the function $y_i = 2^{\sum_j x_{ij}} + \mathcal{N}(0, d^2)$ with $x_{ij} \sim \mathcal{U}[0, 2]$. Fifty simulations were run with 12000 training and 3000 testing points. Only the first 750 partitions of the paths are displayed in order to make the MSE of the earlier IRP iterations visually clearer. Scales also differ in order to make the shapes of the curves clear.

5.2 Modeling MPG of Automobiles

We next illustrate the performance of IRP when predicting the miles-per-gallon of a list of 392 automobiles manufactured between 1970 and 1982 using seven variables (Frank & Asuncion 2010). Seven regressions are performed in dimensions one through seven, where the variables chosen are from the following order: origin (American, European, or Japanese), model year, number of cylinders, acceleration, displacement, horsepower, and weight. The order of the variables was determined in order of the magnitude of coefficients from a least squares linear regression on all variables. Origin, surprisingly, had the largest magnitude, and

Model 1: $y_i = (\prod_j x_{ij}) + \mathcal{N}(0, d^2)$ with $x_{ij} \sim \mathcal{U}[0, 3]$

Number of Variables	IRP Min MSE	Isotonic Regression MSE	LS MSE	IRP Min Path	IRP Path Length
2	**4.06 (± 0.04)**	**4.06 (± 0.04)**	4.54 (± 0.03)	44	437
4	**21.83 (± 0.26)**	**22.37 (± 0.37)**	36.88 (± 0.37)	18	3711
6	391.74 (± 9.52)	866.28 (± 56.18)	385.06 (± 11.45)	3	7685
8	5811.56 (± 312.28)	39088.71 (± 6248.82)	4276.90 (± 287.53)	2	10153

Model 2: $y_i = (\sum_j x_{ij}^2) + \mathcal{N}(0, 4d^2)$ with $x_{ij} \sim \mathcal{U}[0, 5]$

Number of Variables	IRP Min MSE	Isotonic Regression MSE	LS MSE	IRP Min Path	IRP Path Length
2	**16.50 (± 0.13)**	**16.50 (± 0.13)**	23.03 (± 0.13)	625	999
4	**74.68 (± 0.58)**	**75.13 (± 0.61)**	77.94 (± 0.50)	28	4861
6	203.60 (± 1.77)	214.47 (± 2.18)	165.70 (± 1.15)	8	8520
8	596.05 (± 5.42)	655.75 (± 6.15)	285.12 (± 1.81)	4	10604

Model 3: $y_i = 2^{\sum_j x_{ij}} + \mathcal{N}(0, d^2)$ with $x_{ij} \sim \mathcal{U}[0, 2]$

Number of Variables	IRP Min MSE	Isotonic Regression MSE	LS MSE	IRP Min Path	IRP Path Length
2	**4.10 (± 0.02)**	**4.10 (± 0.02)**	4.73 (± 0.04)	62	628
4	**44.47 (± 1.05)**	**49.90 (± 2.01)**	101.99 (± 1.78)	19	7558
6	7847.71 (± 198.92)	16876.51 (± 1055.26)	5008.07 (± 170.40)	4	11787

Model 4: $y_i = (\sum_j x_{ij})^{1/4} + \mathcal{N}(0, d^2/10)$ with $x_{ij} \in \{0, 1, 2\}$ with probabilities $\{.7, .2, .1\}$

Number of Variables	IRP Min MSE	Isotonic Regression MSE	LS MSE	IRP Min Path	IRP Path Length
2	**0.40 (± 0.00)**	**0.40 (± 0.00)**	0.46 (± 0.00)	8	8
4	**1.61 (± 0.01)**	**1.61 (± 0.01)**	1.67 (± 0.01)	9	30
6	3.64 (± 0.03)	3.65 (± 0.03)	3.69 (± 0.03)	4	85
8	6.49 (± 0.04)	6.53 (± 0.04)	6.45 (± 0.05)	5	267

Model 5: $y_i = (\prod_j x_{ij}) + \mathcal{N}(0, d^2)$ with $x_{ij} \in \{0, 1, 2\}$ with probabilities $\{1/3, 1/3, 1/3\}$

Number of Variables	IRP Min MSE	Isotonic Regression MSE	LS MSE	IRP Min Path	IRP Path Length
2	**3.99 (± 0.03)**	**3.99 (± 0.03)**	4.44 (± 0.03)	4	5
4	**15.99 (± 0.12)**	**16.01 (± 0.12)**	20.00 (± 0.16)	6	25
6	**36.44 (± 0.29)**	**36.44 (± 0.29)**	52.91 (± 0.79)	87	103
8	**68.87 (± 0.70)**	**68.88 (± 0.71)**	118.84 (± 4.59)	67	430

Table 1: Statistics for simulations generated by the three different models as labeled above. A path of mean squared errors for each model along the regularization path was computed. *IRP Min MSE* refers to the minimum MSE along these paths. *IRP Min Path* is the number of partitions made to generate the minimum MSE and IRP Path Length is the number of partitions in the global isotonic solution. *LS MSE* is the MSE from using least squares regressions. Bolded MSE values for IRP and isotonic regression indicate that they are lower than the MSE of the least squares regression with 95% confidence.

in giving discrete variables 1,2,3 to the respective origins, there actually is a monotonic trend in origin (i.e., American cars are least fuel efficient, followed by European cars, with the Japanese being most efficient). While we include origin as a variable here, we note that similar performance for IRP was achieved without origin in an independent experiment.

Since the data set is rather small, we perform leave-one-out cross-validation (i.e. the data is divided into training and testing sets of 391 and 1 instances, respectively, so that each instance is used out-of-sample once). Table 2 displays certain statistics on the IRP, and isotonic regression, performance as well as a comparison to the results of a least squares regression. Figure 6 displays MSE on out-of-sample data for IRP over the regularized path of isotonic solutions for a regression with six variables, exemplifying that overfitting occurs after 15 iterations of IRP (seen by the U-shaped curve with minimum at 15 iterations).

Number of Variables	IRP Min MSE	Isotonic Regression MSE	LS MSE	IRP Min Path	IRP Path Length
1	41.79 ± 5.44	42.20 ± 5.39	41.77 ± 6.05	9	17
2	$\mathbf{24.13 \pm 3.53}$	$\mathbf{24.51 \pm 3.50}$	27.41 ± 3.74	7	26
3	$\mathbf{13.88 \pm 2.55}$	$\mathbf{14.12 \pm 2.39}$	16.17 ± 2.90	9	37
4	14.65 ± 2.75	15.30 ± 2.54	16.32 ± 2.93	7	62
5	$\mathbf{11.01 \pm 2.39}$	$\mathbf{11.26 \pm 2.28}$	15.35 ± 2.94	15	109
6	$\mathbf{10.77 \pm 2.37}$	11.34 ± 2.13	14.23 ± 2.70	15	114
7	10.84 ± 2.36	11.28 ± 2.12	11.37 ± 2.09	8	128

Table 2: Statistics for auto mpg data. Miles-per-gallon is regressed on a seven potential variables: origin (American, European, or Japanese), model year, number of cylinders, acceleration, displacement, horsepower, and weight. Row k uses the first k variables from this list in the regression. A path of mean squared errors for each model along the regularization path was computed. Bold demonstrates statistical significance of either IRP or isotonic regression over a least squares regression with 95% confidence, as determined by a paired t-test using 392 observed squared losses obtained from leave-one-out cross-validation.

6 Discussion and extensions

The IRP algorithm offers solutions to both the statistical and computational difficulties of isotonic regression. Algorithmically, IRP solves (2) as a sequence of easier binary partitioning problems that are efficiently solved using network flow algorithms. From the statistical perspective, IRP generates a path of isotonic models, each defining a partitioning of the space $\mathcal{X}$ into isotonic regions. The averages of observations in these regions comply with the isotonicity constraints (Theorem 4). In this view, IRP provides isotonic solutions along its path that are regularized versions of the globally optimal isotonic regression solution.

Our discussion so far has focused on using the sum of squares loss function in (2) for fitting "standard" isotonic regression subject to squared error loss. A well known result of Barlow & Brunk (1972) implies that the solution of a whole variety of loss functions subject to isotonicity constraints can be obtained by solving standard isotonic regression, as long as the loss can be written as minimizing $\sum_{i=1}^{n} w_i(\Psi(z_i) - x_i z_i)^2$ in $z \in \mathbf{R}^n$ for some convex differentiable Ψ and some data-dependent values x and weights w.

We first show how this result facilitates a connection between IRP and the well-known work of Maxwell & Muckstadt (1985) (and similarly Roundy (1986)), who solved an operations research problem (related to scheduling reorder intervals for a production system) by reducing it to the optimization of a convex objective

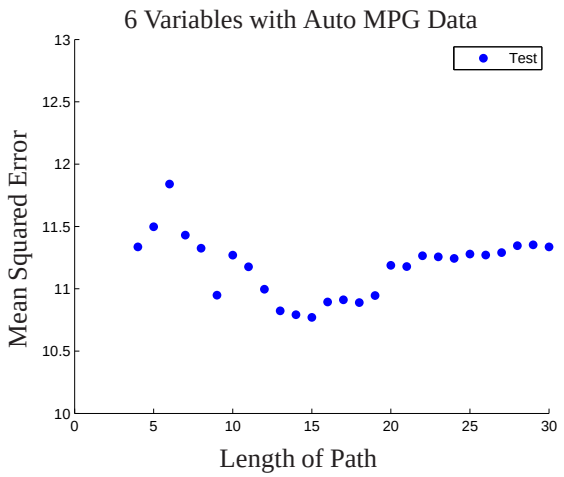

Figure 6: Mean squared error for auto data with six variables using IRP illustrates that overfitting occurs after 15 iterations of IRP, i.e. MSE decreases until 15 partitions are made, after which MSE begins to increase. The figure displays only the first 30 partitions so that the U-shape is clear.

subject to isotonicity constraints. In our notation, their objective (i.e. loss function) is $\sum_{i=1}^{n}(c_i/\hat{y}_i + b_i\hat{y}_i)$, where c_i, b_i are data-dependent nonnegative constants determined by their problem formulation. To apply the theory of Barlow & Brunk (1972), we reformulate their problem as minimizing $\sum_{i=1}^{n} c_i(z_i - (-b_i/c_i))^2$ in $z \in \mathbf{R}^n$, i.e. a standard weighted isotonic regression, and recovering $\hat{y}_i^* = \sqrt{-z_i^*}$ (note that the isotonic regression fits nonpositive observations $-b_i/c_i$). Indeed, the algorithm of Maxwell & Muckstadt (1985) is completely equivalent to applying IRP on this modified problem! It should be emphasized, however, that Maxwell & Muckstadt (1985) were interested in this algorithm purely as a means to reach the globally optimal solution, and were uninterested in statistical considerations which led us to consider intermediate IRP solutions as regularized isotonic models of independent interest. Spouge et al. (2003) also used Maxwell & Muckstadt (1985) to inspire the partitioning algorithm for the standard isotonic regression problem, however they do not make the connection using Barlow & Brunk (1972), and also have no statistical interests in mind.

The results of Barlow & Brunk (1972) also imply that many other loss functions subject to isotonicity constraints can optimally be solved via a reformulation to a problem of the form (2). For instance, in the case of a binary response $y \in \{0, 1\}$, it may be desirable to fit isotonic models by minimizing the in-sample logistic log likelihood rather than the sum of squares (Bacchetti 1989, Auh & Sampson 2006):

$$\min\{\sum_{i=1}^{n} y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i) : \hat{p}_i \leq \hat{p}_j \quad \forall (i,j) \in \mathcal{I}, 0 \leq \hat{p}_i \leq 1 \quad \forall i\}. \tag{11}$$

Applying the transformation of Barlow & Brunk (1972), the solution to (11) turns out to be identical to solving the squared loss problem (2) with the values $y_i \in \{0, 1\}$. Naturally, any problem that can be reformulated as (2) can be solved using the IRP algorithm, and we plan to investigate the applicability of the resulting regularization algorithms in future work.

As our analysis and experiments have demonstrated, computation is not a significant concern with IRP, at least for moderate to large data sizes. However, overfitting is still a major concern as dimensionality

grows. As demonstrated in Sections 4 and 5, while IRP offers partial protection from overfitting through its regularization behavior, even the first step in the IRP path could already suffer from high variance in dimension as low as six. A key question pertains to identifying factors affecting this overfitting behavior, specifically characterizing situations in which the initial IRP iterations are less prone to overfitting.

7 Appendix

Proposition 1:

Proof. We first rewrite both the IRP partition problem (3) and the maximal between-group variance partition problem (4). Assume $V = A \cup B$ and $A \cap B = \{\}$. Then it is easy to show $|V|\overline{y}_V = |A|\overline{y}_A + |B|\overline{y}_B$ which gives $(\overline{y}_A - \overline{y}_V) = -|B|(\overline{y}_B - \overline{y}_V)/|A|$. The objective function to (3) can be written $|B|(\overline{y}_B - \overline{y}_V) - |A|(\overline{y}_A - \overline{y}_V)$ and using the previous relationship can again be rewritten $2|B|(\overline{y}_B - \overline{y}_V)$. An obvious property of the optimal IRP cut is that $\overline{y}_B \geq \overline{y}_V$. If we add this as a redundant constraint to the IRP partition (3), then we can find the same optimal partition by maximizing the square of the objective, i.e. maximize $4|B|^2(\overline{y}_B - \overline{y}_V)^2$ subject to the appropriate constraints. The objective of the between-group variance partition (4) can be rewritten using the above relationship as $(|B| + |B|^2/|A|)(\overline{y}_B - \overline{y}_V)^2$. Then denoting the IRP and maximal between-group variance objectives by $g^*(A, B)$ and $\tilde{g}(A, B)$ respectively, we have $g^*(A, B) = 4|A||B|\tilde{g}(A, B)/n$ since $|A| + |B| = n$ is constant. Eliminating the constant $4/n$ gives the first result.

In order to prove the second statement, notice that optimality of (3) and (4) gives $|A^*||B^*|\tilde{g}(A^*, B^*) \geq |A^*||B^*|\tilde{g}(\tilde{A}, \tilde{B})$ and $\tilde{g}(\tilde{A}, \tilde{B}) \geq \tilde{g}(A^*, B^*)$ which implies $|A^*||B^*| \geq |\tilde{A}||\tilde{B}|$. This along with the relation

$$(|A^*| + |B^*|)^2 = |A^*|^2 + 2|A^*||B^*| + |B^*|^2 = |\tilde{A}|^2 + 2|\tilde{A}||\tilde{B}| + |\tilde{B}|^2 = (|\tilde{A}| + |\tilde{B}|)^2$$

gives $|A^*|^2 + |B^*|^2 \leq |\tilde{A}|^2 + |\tilde{B}|^2$. We use this to get the relation

$$\begin{aligned}&(|A^*| - |B^*|)^2 = |A^*|^2 - 2|A^*||B^*| + |B^*|^2 \\ &\leq |\tilde{A}|^2 - 2|A^*||B^*| + |\tilde{B}|^2 \leq |\tilde{A}|^2 - 2|\tilde{A}||\tilde{B}| + |\tilde{B}|^2 = (|\tilde{A}| - |\tilde{B}|)^2\end{aligned}$$

which gives the second result of the proposition. ■

Theorem 2:

Proof. Divide the blocks in V into three subsets:

1. $\mathcal{L}$: union of all blocks in V that are "below" the algorithm cut.
2. $\mathcal{U}$: union of all blocks in V that are "above" the algorithm cut.
3. $\mathcal{M}$: union of K blocks in V that get broken by the cut (note that blocks in $\mathcal{M}$ may be separated by blocks in $\mathcal{L}$ or $\mathcal{U}$).

Define M_1 (M_K) to be the minorant (majorant) block in $\mathcal{M}$. For each M_k define M_k^L (M_k^U) as the groups in M_k below (above) the algorithm cut. Define $A_K^L \subseteq \mathcal{L}$ ($A_1^U \subseteq \mathcal{U}$) as the union of blocks along the algorithm cut such that $A_K^L \succ M_K^L$ ($A_1^U \prec M_1^U$). Refer to Figure 7 for an example of these definitions where $A_1^U = A_1^L = A_K^U = A_K^L = \{\}$ for simplicity.

We use the above definitions and assumptions to state the following two consequences that cause a contradiction:

I $\overline{y}_{M_1} < \overline{y}_{M_K}$ by optimality (i.e. according to KKT conditions) and isotonicity.

II $\overline{y}_{M_1} > \overline{y}_V$ and $\overline{y}_{M_K} < \overline{y}_V$. This is proven below.

(II) implies $\overline{y}_{M_1} > \overline{y}_{M_K}$ which contradicts (I) and we are left to prove (II). Optimality of blocks M_1 and M_K gives

(a) $\overline{y}_{M_1^L} > \overline{y}_{M_1^U}$

(b) $\overline{y}_{M_K^L} > \overline{y}_{M_K^U}$.

The proof for $\overline{y}_{M_1} > \overline{y}_V$ is as follows with two cases:

1. $A_1^U = \{\}$: $\overline{y}_{M_1^U} > \overline{y}_V$ because using the algorithm cut in (5), we have

$$\sum_{i \in M_1^U} (y_i - \overline{y}_V) > 0 \Rightarrow \sum_{i \in M_1^U} y_i > |M_1^U| \overline{y}_V \Rightarrow \overline{y}_{M_1^U} > \overline{y}_V.$$

   The first inequality is true about the cut because there exist no block below M_1^U to affect isotonicity. Then using (a), we get

$$\overline{y}_{M_1^L} > \overline{y}_{M_1^U} > \overline{y}_V \Rightarrow \overline{y}_{M_1} > \overline{y}_V$$

2. $A_1^U \neq \{\}$: $\overline{y}_{M_1} > \overline{y}_{A_1^U} > \overline{y}_V$. The first inequality is due to optimality and the second is again because the algorithm cut in (5) gives

$$\sum_{i \in A_1^U} (y_i - \overline{y}_V) > 0 \Rightarrow \sum_{i \in A_1^U} y_i > |A_1^U| \overline{y}_V \Rightarrow \overline{y}_{A_1^U} > \overline{y}_V,$$

   which again is possible because no block exists below A_1^U to affect isotonicity.

The proof for $\overline{y}_{M_K} < \overline{y}_V$ is a similar argument and hence gives (II). The case $K = 1$ is also trivially covered by the above arguments. We conclude that the algorithm cannot cut any block. ■

The following remark is necessary for completeness of the proof of Theorem 2.

Remark 6 *The case of two connected optimal groups having equal means need not be discussed in Theorem 2. In this event, the optimal solution to isotonic regression in not unique. It is trivial that M_1 would not have been split by Algorithm 1 if $\overline{y}_{M_1^L} = \overline{y}_{M_1^U} \neq \overline{y}_V$. Otherwise, consider the case $\overline{y}_{M_1^L} = \overline{y}_{M_1^U} = \overline{y}_V$ and assume M_1 is a block broken by the cut in V. M_1^L and M_1^U are also possible blocks whereby $M_1^L \in \mathcal{L}$ and $M_1^U \in \mathcal{U}$, and hence $M_1 = M_1^L \cup M_1^U \notin \mathcal{M}$. The same remarks apply to M_K. Thus, the proof still holds if there are multiple isotonic solutions.*

Remark 7 *The case of multiple observations at the same coordinates can be disregarded. To see this, let J be a set of nodes with the same coordinates. From the constraints, $y_i = y_j, \forall i, j \in J$ and thus the number of observations can be reduced and all observations in J fit to the same value $\hat{y}$. Then*

$$\sum_{j \in J} (\hat{y} - y_j)^2 = |J| \mid (\hat{y} - \overline{y}_J)^2 + \sum_{j \in J} y_j^2 - \overline{y}_J^2$$

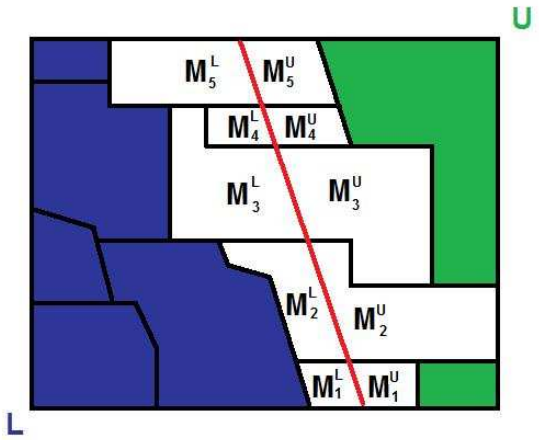

Figure 7: Illustration of proof of Theorem 2. Black lines separate blocks. The diagonal red line through the center demonstrates a cut of Algorithm 1. $\mathcal{L}$ is the union of blue blocks below the cut and $\mathcal{U}$ is the union of green blocks above the cut. White blocks are blocks that are potentially split by Algorithm 1. These blocks are split into $M_1^L, \ldots, M_5^L$ below the cut and $M_1^U, \ldots, M_5^U$ above the cut. In the proof, $M_i = M_i^L \cup M_i^U \ \forall i = 1 \ldots 5$. The proof shows, for example, that if the algorithm splits M_1 into M_1^L and M_1^U according to the defined cut in (5), then there must be no isotonicity violation when creating blocks from M_1^L and M_1^U. However, since M_1 is assumed to be a block, there must exist an isotonicity violation between M_1^L and M_1^U, providing a contradiction.

so that the sum of squared differences over J can be reduced to be a single weighted squared difference. Problem (2) becomes the weighted isotonic regression problem

$$\min \{\sum_{i=1}^{n} w_i(\hat{y}_i - y_i)^2 : \hat{y}_i \leq \hat{y}_j \quad \forall (i,j) \in \mathcal{I}\}, \tag{12}$$

for which the KKT conditions imply that observations are again divided into k groups where the fits in each group take the weighted group mean $\overline{y}_V^w = \sum_{i \in V} (w_i y_i) / \sum_{i \in V} w_i$ rather than the group mean. The optimal cut problem (5) changes to have $z_i = w_i(y_i - \overline{y}_V^w)$ and the above results on IRP generalize easily noting that now the weighted algorithm cut implies $\overline{y}_A^w > \overline{y}_V^w$ for a group A on the upper side of the cut such that no group exists below A that could affect isotonicity.

Proposition 5:

Proof. Any final partition can be represented by a simple tree. Consider level k of the tree. Let $p_k \geq .5$ be the greatest p over levels $1, \ldots, k-1$ such that a partition of group size n_k into two groups of size pn_k and $(1-p)n_k$ where n_k is the corresponding size of the partitioned group. Denote by L_k the largest group partitioned at iteration k whose size can be bounded by $|L_k| \leq np_k^k$. We next note that the complexity of solving a problem with n observations is higher than solving 2 problems with pn and $(1-p)n$ observations. Indeed, $n^3 = pn^3 + (1-p)n^3 > p^2n^3 + (1-p)^2n^3$. Thus, we assume that at iteration k, we solve only problems of the largest possible size (rather than several problems of small size). The number of groups at iteration k can also be bounded by $n/|L_k|$. Denote by $T_{p_k}(k)$ the complexity of partitioning all groups at

level k. Then

$$T_{p_k}(k) \leq O(\frac{n}{|L|}|L|^3) = O(n|L|^2) \leq O(n(np_k^k)^2) = O(n^3)p_k^{2k}.$$

Then denote by K the total number of levels in the partition tree. We have

$$\sum_{k=1}^{K} T_{p_k}(k) \leq \sum_{k=1}^{K} O(n^3)p_{\max}^{2k} \leq \sum_{k=1}^{\infty} O(n^3)p_{\max}^{2k} = O(n^3)\frac{1}{1-p_{\max}^2}.$$

■

8 Acknowledgements

The authors are grateful to Quentin Stout for drawing our attention to additional relevant references.

References

Auh, S. & Sampson, A. R. (2006), 'Isotonic logistic discrimination', *Biometrika* **93**(4), 961–972.

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

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

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

Boyd, S. & Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. (1984), *Classification and Regression Trees*, Chapman and Hall/CRC.

Chandrasekaran, R., Ryu, Y., Jacob, V. & Hong, S. (2005), 'Isotonic separation', *INFORMS Journal on Computing* **17**(4), 462–474.

Cordell, H. J. (2009), 'Detecting genegene interactions that underlie human diseases', *Nature Reviews Genetics* **10**, 392–404.

de Leeuw, J., Hornik, K. & Mair, P. (2009), 'Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods'. UC Los Angeles: Department of Statistics, UCLA. Retrieved from: http://cran.r-project.org/web/packages/isotone/vignettes/isotone.pdf.

Dykstra, R. L. & Robertson, T. (1982), 'An algorithm for isotonic regression for two or more independent variables', *Annals of Statistics* **10**(3), 708–716.

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

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. & Nadeau, J. H. (2010), 'Missing heritability and strategies for finding the underlying causes of complex disease', *Nature Reviews Genetics* **11**, 446–450.

Frank, A. & Asuncion, A. (2010), 'UCI machine learning repository'. Auto MPG Data Set available at http://archive.ics.uci.edu/ml.

Galil, Z. & Naamad, A. (1980), 'An o(EV$\log^2$V) algorithm for the maximal flow problem', *Journal of the Computer and System Sciences* **21**, 203–217.

Goldstein, D. B. (2009), 'Common genetic variation and human traits', *NEJM* **360**, 1696–1698.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer.

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

Kruskal, J. (1964), 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika* **29**(1).

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

Mani, R., St.Onge, R. P., Hartman, J. L., Giaever, G. & Roth, F. P. (2007), 'Defining genetic interaction', *PNAS* **105**(9), 3461–3466.

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

Meyer, M. & Woodroofe, M. (2000), 'On the degrees of freedom in shape-restricted regression', *Annals of Statistics* **28**(4), 1083–1104.

Monteiro, R. & Adler, I. (1989), 'Interior path following primal-dual algorithms. part II: Convex quadratic programming', *Mathematical Programming* **44**, 43–66.

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

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

Roth, F. P., Lipshitz, H. D. & Andrews, B. J. (2009), 'Q&a: Epistasis', *Journal of Biology* **8**(4). Article 35.

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.

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

Shao, H., Burrage, L. C., Sinasac, D. S., Hill, A. E., Ernest, S. R., O'Brien, W., Courtland, H.-W., Jepsen, K. J., Kirby, A., Kulbokas, E. J., Daly, M. J., Bromang, K. W., Lander, E. S. & Nadeau, J. H. (2008), 'Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis', *PNAS* (50), 11910–19914.

Sleator, D. & Tarjan, R. E. (1983), 'A data structure for dynamic trees', *Journal of Computer and System Sciences* **26**(3), 362–391.

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

Stein, C. M. (1981), 'Estimation of the mean of a multivariate normal distribution', *The Annals of Statistics* **9**(6), 1131–1151.

Stout, Q. (2010), 'An approach to computing multidimensional isotonic regressions'. Submitted. Available at: http://www.eecs.umich.edu/ qstout/pap/MultidimIsoReg.pdf.

Ye, J. (1998), 'On measuring and correcting the effects of data mining and model selection', *Journal of the American Statistical Association* **93**(441), 120–131.