Efficient Learning of Generalized Linear and Single Index Models with Isotonic Regression

Sham M. Kakade
Microsoft Research and Wharton, U Penn
skakade@microsoft.com

Adam Tauman Kalai
Microsoft Research
adum@microsoft.com

Varun Kanade
SEAS, Harvard University
vkanade@fas.harvard.edu

Ohad Shamir
Microsoft Research
ohadsh@microsoft.com

Abstract

Generalized Linear Models (GLMs) and Single Index Models (SIMs) provide powerful generalizations of linear regression, where the target variable is assumed to be a (possibly unknown) 1-dimensional function of a linear predictor. In general, these problems entail non-convex estimation procedures, and, in practice, iterative local search heuristics are often used. Kalai and Sastry (2009) provided the first provably efficient method, the *Isotron* algorithm, for learning SIMs and GLMs, under the assumption that the data is in fact generated under a GLM and under certain monotonicity and Lipschitz (bounded slope) constraints. The Isotron algorithm interleaves steps of perceptron-like updates with isotonic regression (fitting a one-dimensional non-decreasing function). However, to obtain provable performance, the method requires a fresh sample every iteration. In this paper, we provide algorithms for learning GLMs and SIMs, which are both computationally and statistically efficient. We modify the isotonic regression step in Isotron to fit a Lipschitz monotonic function, and also provide an efficient $O(n \log(n))$ algorithm for this step, improving upon the previous $O(n^2)$ algorithm. We provide a brief empirical study, demonstrating the feasibility of our algorithms in practice.

1 Introduction

The oft used linear regression paradigm models a dependent variable Y as a linear function of a vector-valued independent variable X. Namely, for some vector w, we assume that $\mathbb{E}[Y|X] = w \cdot X$. Generalized linear models (GLMs) provide a flexible extension of linear regression, by assuming that the dependent variable Y is of the form, $\mathbb{E}[Y|X] = u(w \cdot X)$; u is referred to as the inverse link function or transfer function (see [1] for a review). Generalized linear models include commonly used regression techniques such as logistic regression, where $u(z) = 1/(1 + e^{-z})$ is the logistic function. The class of perceptrons also falls in this category, where u is a simple piecewise linear function of the form _/‾, with the slope of the middle piece being the inverse of the margin.

In the case of linear regression, the least-squares method is an highly efficient procedure for parameter estimation. Unfortunately, in the case of GLMs, even in the setting when u is known, the problem of fitting a model that minimizes squared error is typically not convex. We are not aware of any classical estimation procedure for GLMs which is both *computationally* and *statistically* efficient, and with provable guarantees. The standard procedure is iteratively reweighted least squares, based on Newton-Raphson (see [1]).

The case when *both* u and w are unknown (sometimes referred to as Single Index Models (SIMs)), involves the more challenging (and practically relevant) question of jointly estimating u and w,

where u may come from a large non-parametric family such as all monotonic functions. There are two questions here: 1) What statistical rate is achievable for simultaneous estimation of u and w? 2) Is there a *computationally efficient* algorithm for this joint estimation? With regards to the former, under mild Lipschitz-continuity restrictions on u, it is possible to characterize the effectiveness of an (appropriately constrained) joint empirical risk minimization procedure. This suggests that, from a purely statistical viewpoint, it may be worthwhile to attempt jointly optimizing u and w on empirical data.

However, the issue of *computationally efficiently* estimating both u and w (and still achieving a good statistical rate) is more delicate, and is the focus of this work. We note that this is not a trivial problem: in general, the joint estimation problem is highly non-convex, and despite a significant body of literature on the problem, existing methods are usually based on heuristics, which are not guaranteed to converge to a global optimum (see for instance [2, 3, 4, 5, 6]).

The Isotron algorithm of Kalai and Sastry [7] provides the first provably efficient method for learning GLMs and SIMs, under the common assumption that u is monotonic and Lipschitz, and assuming that the data corresponds to the model.[1] The sample and computational complexity of this algorithm is polynomial, and the sample complexity does not explicitly depend on the dimension. The algorithm is a variant of the "gradient-like" perceptron algorithm, where apart from the perceptron-like updates, an isotonic regression procedure is performed on the linear predictions using the Pool Adjacent Violators (PAV) algorithm, on every iteration.

While the Isotron algorithm is appealing due to its ease of implementation (it has no parameters other than the number of iterations to run) and theoretical guarantees (it works for any u, w), there is one principal drawback. It is a batch algorithm, but the analysis given requires the algorithm to be run on *fresh samples* each batch. In fact, as we show in experiments, this is not just an artifact of the analysis – if the algorithm loops over the same data in each update step, it really does overfit in very high dimensions (such as when the number of dimensions exceeds the number of examples).

Our Contributions: We show that the overfitting problem in Isotron stems from the fact that although it uses a slope (Lipschitz) condition as an assumption in the analysis, it does not constrain the output hypothesis to be of this form. To address this issue, we introduce the SLISOTRON algorithm (pronounced slice-o-tron, combining slope and Isotron). The algorithm replaces the isotonic regression step of the Isotron by finding the best non-decreasing function *with a bounded Lipschitz parameter* - this constraint plays here a similar role as the margin in classification algorithms. We also note SLISOTRON (like Isotron) has a significant advantage over standard regression techniques, since it does not require knowing the transfer function. Our two main contributions are:

1. We show that the new algorithm, like Isotron, has theoretical guarantees, and significant new analysis is required for this step.

2. We provide an efficient $O(n \log(n))$ time algorithm for finding the best non-decreasing function with a bounded Lipschitz parameter, improving on the previous $O(n^2)$ algorithm [10]. This makes SLISOTRON practical even on large datasets.

We begin with a simple perceptron-like algorithm for fitting GLMs, with a *known* transfer function u which is monotone and Lipschitz. Somewhat surprisingly, prior to this work (and Isotron [7]) a computationally efficient procedure that guarantees to learn GLMs was not known. Section 4 contains the more challenging SLISOTRON algorithm and also the efficient $O(n \log(n))$ algorithm for Lipschitz isotonic regression. We conclude with a brief empirical analysis.

2 Setting

We assume the data (x, y) are sampled i.i.d. from a distribution supported on $\mathbb{B}_d \times [0, 1]$, where $\mathbb{B}_d = \{x \in \mathbb{R}^d \ : \ \|x\| \leq 1\}$ is the unit ball in d-dimensional Euclidean space. Our algorithms and

[1] In the more challenging agnostic setting, the data is not required to be distributed according to a true u and w, but it is required to find the best u, w which minimize the empirical squared error. Similar to observations of Kalai et al. [8], it is straightforward to show that this problem is likely to be computationally intractable in the agnostic setting. In particular, it is at least as hard as the problem of "learning parity with noise," whose hardness has been used as the basis for designing multiple cryptographic systems. Shalev-Shwartz et al. [9] present a kernel-based algorithm for learning certain types of GLMs and SIMs in the agnostic setting. However, their worst-case guarantees are exponential in the norm of w (or equivalently the Lipschitz parameter).

Algorithm 1 GLM-TRON

Input: data $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{R}^d \times [0,1]$, $u : \mathbb{R} \to [0,1]$, held-out data $\langle (x_{m+j}, y_{m+j}) \rangle_{j=1}^s$

$w^1 := 0$;

for $t = 1, 2, \ldots$ **do**

$h^t(x) := u(w^t \cdot x)$;

$w^{t+1} := w^t + \frac{1}{m} \sum_{i=1}^m (y_i - u(w^t \cdot x_i)) x_i$;

end for

Output: $\arg\min_{h^t} \sum_{j=1}^s (h^t(x_{m+j}) - y_{m+j})^2$

analysis also apply to the case where $\mathbb{B}_d$ is the unit ball in some high (or infinite)-dimensional kernel feature space. We assume there is a fixed vector w, such that $\|w\| \leq W$, and a non-decreasing 1-Lipschitz function $u : \mathbb{R} \to [0,1]$, such that $\mathbb{E}[y|x] = u(w \cdot x)$ for all x. The restriction that u is 1-Lipschitz is without loss of generality, since the norm of w is arbitrary (an equivalent restriction is that $\|w\| = 1$ and that u is W-Lipschitz for an arbitrary W).

Our focus is on approximating the regression function well, as measured by the squared loss. For a real valued function $h : \mathbb{B}_d \to [0,1]$, define

$$\mathrm{err}(h) = \mathbb{E}_{(x,y)} \left[(h(x) - y)^2 \right]$$
$$\varepsilon(h) = \mathrm{err}(h) - \mathrm{err}(E[y|x]) \ = \mathbb{E}_{(x,y)} \left[(h(x) - u(w \cdot x))^2 \right]$$

$\mathrm{err}(h)$ measures the error of h, and $\varepsilon(h)$ measures the excess error of h compared to the Bayes-optimal predictor $x \mapsto u(w \cdot x)$. Our goal is to find h such that $\varepsilon(h)$ (equivalently, $\mathrm{err}(h)$) is as small as possible.

In addition, we define the empirical counterparts $\widehat{\mathrm{err}}(h)$, $\hat{\varepsilon}(h)$, based on a sample $(x_1, y_1), \ldots, (x_m, y_m)$, to be

$$\widehat{\mathrm{err}}(h) \ = \frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)^2; \qquad \hat{\varepsilon}(h) \ = \frac{1}{m} \sum_{i=1}^m (h(x_i) - u(w \cdot x_i))^2.$$

Note that $\hat{\varepsilon}$ is the standard *fixed design error* (as this error conditions on the observed x's).

Our algorithms work by iteratively constructing hypotheses h^t of the form $h^t(x) = u^t(w^t \cdot x)$, where u^t is a non-decreasing, 1-Lipschitz function, and w^t is a linear predictor. The algorithmic analysis provides conditions under which $\hat{\varepsilon}(h^t)$ is small, and using statistical arguments, one can guarantee that $\varepsilon(h^t)$ would be small as well.

3 The GLM-TRON algorithm

We begin with the simpler case, where the transfer function u is assumed to be known (e.g. a sigmoid), and the problem is estimating w properly. We present a simple, parameter-free, perceptron-like algorithm, GLM-TRON (Alg. 1), which efficiently finds a close-to-optimal predictor. We note that the algorithm works for arbitrary non-decreasing, Lipschitz functions u, and thus covers most generalized linear models. We refer the reader to the pseudo-code in Algorithm 1 for some of the notation used in this section.

To analyze the performance of the algorithm, we show that if we run the algorithm for sufficiently many iterations, one of the predictors h^t obtained must be nearly-optimal, compared to the Bayes-optimal predictor.

Theorem 1. *Suppose $(x_1, y_1), \ldots, (x_m, y_m)$ are drawn independently from a distribution supported on $\mathbb{B}_d \times [0,1]$, such that $\mathbb{E}[y|x] = u(w \cdot x)$, where $\|w\| \leq W$, and $u : \mathbb{R} \to [0,1]$ is a known non-decreasing 1-Lipschitz function. Then for any $\delta \in (0,1)$, the following holds with probability at least $1 - \delta$: there exists some iteration $t < O(W\sqrt{m/\log(1/\delta)})$ of* GLM-TRON *such that the hypothesis $h^t(x) = u(w^t \cdot x)$ satisfies*

$$\max\{\hat{\varepsilon}(h^t), \varepsilon(h^t)\} \leq O\left(\sqrt{\frac{W^2 \log(m/\delta)}{m}} \right).$$

Algorithm 2 SLISOTRON

Input: data $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{R}^d \times [0,1]$, held-out data $\langle (x_{m+j}, y_{m+j}) \rangle_{j=1}^s$
$w^1 := 0$;
for $t = 1, 2, \ldots$ **do**
  $u^t := \mathsf{LIR}\left((w^t \cdot x_1, y_1), \ldots, (w^t \cdot x_m, y_m)\right)$ // Fit 1-d function along w^t
  $w^{t+1} := w^t + \frac{1}{m} \sum_{i=1}^m (y_i - u^t(w^t \cdot x_i)) x_i$
end for
Output: $\arg\min_{h^t} \sum_{j=1}^s (h^t(x_{m+j}) - y_{m+j})^2$

In particular, the theorem implies that some h^t has small enough $\varepsilon(h^t)$. Since $\varepsilon(h^t)$ equals $\mathrm{err}(h^t)$ up to a constant, we can easily find an appropriate h^t by picking the one that has least $\widehat{\mathrm{err}}(h^t)$ on a held-out set.

The main idea of the proof is showing that at each iteration, if $\hat{\varepsilon}(h^t)$ is not small, then the squared distance $\left\|w^{t+1} - w\right\|^2$ is substantially smaller than $\|w^t - w\|^2$. Since the squared distance is bounded below by 0, and $\left\|w^0 - w\right\|^2 \leq W^2$, there is an iteration (arrived at within reasonable time) such that the hypothesis h^t at that iteration is highly accurate. Although the algorithm minimizes empirical squared error, we can bound the true error using a uniform convergence argument. The complete proofs are provided in Appendix A.

4 The SLISOTRON algorithm

In this section, we present SLISOTRON (Alg. 2), which is applicable to the harder setting where the transfer function u is unknown, except for it being non-decreasing and 1-Lipschitz. SLISOTRON does have one parameter, the Lipschitz constant; however, in theory we show that this can simply be set to 1. The main difference between SLISOTRON and GLM-TRON is that now the transfer function must also be learned, and the algorithm keeps track of a transfer function u^t which changes from iteration to iteration. The algorithm is inspired by the Isotron algorithm [7], with the main difference being that at each iteration, instead of applying the PAV procedure to fit an arbitrary monotonic function along the direction w^t, we use a different procedure, (Lipschitz Isotonic Regression) LIR, to fit a *Lipschitz* monotonic function, u^t, along w^t. This key difference allows for an analysis that does not require a fresh sample each iteration. We also provide an efficient $O(m \log(m))$ time algorithm for LIR (see Section 4.1), making SLISOTRON an extremely efficient algorithm.

We now turn to the formal theorem about our algorithm. The formal guarantees parallel those of the GLM-TRON algorithm. However, the rates achieved are somewhat worse, due to the additional difficulty of simultaneously estimating both u and w.

Theorem 2. *Suppose $(x_1, y_1), \ldots, (x_m, y_m)$ are drawn independently from a distribution supported on $\mathbb{B}_d \times [0,1]$, such that $\mathbb{E}[y|x] = u(w \cdot x)$, where $\|w\| \leq W$, and $u : \mathbb{R} \to [0,1]$ is an unknown non-decreasing 1-Lipschitz function. Then the following two bounds hold:*

1. *(Dimension-dependent) With probability at least $1 - \delta$, there exists some iteration $t < O\left(\left(\frac{Wm}{d \log(Wm/\delta)}\right)^{1/3}\right)$ of* SLISOTRON *such that*

$$\max\{\hat{\varepsilon}(h^t), \varepsilon(h^t)\} \leq O\left(\left(\frac{dW^2 \log(Wm/\delta)}{m}\right)^{1/3}\right).$$

2. *(Dimension-independent) With probability at least $1 - \delta$, there exists some iteration $t < O\left(\left(\frac{Wm}{\log(m/\delta)}\right)^{1/4}\right)$ of* SLISOTRON *such that*

$$\max\{\hat{\varepsilon}(h^t), \varepsilon(h^t)\} \leq O\left(\left(\frac{W^2 \log(m/\delta)}{m}\right)^{1/4}\right)$$

As in the case of Thm. 1, one can easily find h^t which satisfies the theorem's conditions, by running the SLISOTRON algorithm for sufficiently many iterations, and choosing the hypothesis h^t which minimizes $\widehat{\mathrm{err}}(h^t)$ on a held-out set. The algorithm minimizes empirical error and generalization bounds are obtained using a uniform convergence argument. The proofs are somewhat involved and appear in Appendix B.

4.1 Lipschitz isotonic regression

The SLISOTRON algorithm (Alg. 2) performs Lipschitz Isotonic Regression (LIR) at each iteration. The goal is to find the best fit (least squared error) non-decreasing 1-Lipschitz function that fits the data in one dimension. Let $(z_1, y_1), \ldots (z_m, y_m)$ be such that $z_i \in \mathbb{R}$, $y_i \in [0, 1]$ and $z_1 \leq z_2 \leq \cdots \leq z_m$. The *Lipschitz Isotonic Regression* (LIR) problem is defined as the following quadratic program:

$$\text{Minimize w.r.t } \hat{y}_i \ : \frac{1}{2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{1}$$

subject to:

$$\hat{y}_i \leq \hat{y}_{i+1} \qquad 1 \leq i \leq m-1 \text{ (Monotonicity)} \tag{2}$$

$$\hat{y}_{i+1} - \hat{y}_i \leq (z_{i+1} - z_i) \qquad 1 \leq i \leq m-1 \text{ (Lipschitz)} \tag{3}$$

Once the values $\hat{y}_i$ are obtained at the data points, the actual function can be constructed by interpolating linearly between the data points. Prior to this work, the best known algorithm for this problem wass due to Yeganova and Wilbur [10] and required $O(m^2)$ time for m points. In this work, we present an algorithm that performs the task in $O(m \log(m))$ time. The actual algorithm is fairly complex and relies on designing a clever data structure. We provide a high-level view here; the details are provided in Appendix D.

Algorithm Sketch: We define functions $G_i(\cdot)$, where $G_i(s)$ is the minimum squared loss that can be attained if $\hat{y}_i$ is fixed to be s, and $\hat{y}_{i+1}, \ldots \hat{y}_m$ are then chosen to be the best fit 1-Lipschitz non-decreasing function to the points $(z_i, y_i), \ldots, (z_m, y_m)$. Formally, for $i = 1, \ldots, m$, define the functions,

$$G_i(s) = \min_{\hat{y}_{i+1}, \ldots, \hat{y}_m} \frac{1}{2}(s - y_i)^2 + \frac{1}{2}\sum_{j=i+1}^{m}(\hat{y}_j - y_j)^2 \tag{4}$$

subject to the constraints (where $s = \hat{y}_i$),

$$\hat{y}_j \leq \hat{y}_{j+1} \qquad i \leq j \leq m-1 \text{ (Monotonic)}$$

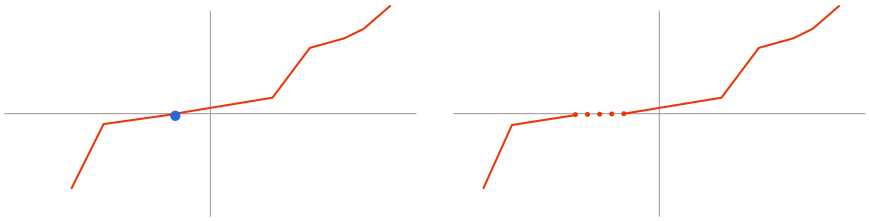
$$\hat{y}_{j+1} - \hat{y}_j \leq z_{j+1} - z_j \qquad i \leq j \leq m-1 \text{ (Lipschitz)}$$

Furthermore, define: $s_i^* = \min_s G_i(s)$. The functions G_i are piecewise quadratic, differentiable everywhere and strictly convex, a fact we prove in full paper [11]. Thus, G_i is minimized at s_i^* and it is strictly increasing on both sides of s_i^*. Note that $G_m(s) = (1/2)(s - y_m)^2$ and hence is piecewise quadratic, differentiable everywhere and strictly convex. Let $\delta_i = z_{i+1} - z_i$. The remaining G_i obey the following recursive relation.

$$G_{i-1}(s) = \frac{1}{2}(s - y_{i-1})^2 + \begin{cases} G_i(s + \delta_{i-1}) & \text{If } s \leq s_i^* - \delta_{i-1} \\ G_i(s_i^*) & \text{If } s_i^* - \delta_{i-1} < s \leq s_i^* \\ G_i(s) & \text{If } s_i^* < s \end{cases} \tag{5}$$

As intuition for the above relation, note that $G_{i-1}(s)$ is obtained fixing $\hat{y}_{i-1} = s$ and then by choosing $\hat{y}_i$ as close to s_i^* (since G_i is strictly increasing on both sides of s_i^*) as possible without violating either the monotonicity or Lipschitz constraints.

The above argument can be immediately translated into an algorithm, if the values s_i^* are known. Since s_1^* minimizes $G_1(s)$, which is the same as the objective of (1), start with $\hat{y}_1 = s_1^*$, and then successively chose values for $\hat{y}_i$ to be as close to s_i^* as possible without violating the Lipschitz or monotonicity constraints. This will produce an assignment for $\hat{y}_i$ which achieves loss equal to $G_1(s_1^*)$ and hence is optimal.



(a) (b)

Figure 1: (a) Finding the zero of G'_i. (b) Update step to transform representation of G'_i to G'_{i-1}

The harder part of the algorithm is finding the values s^*_i. Notice that G'_i are all piecewise linear, continuous and strictly increasing, and obey a similar recursive relation ($G'_m(s) = s - y_m$):

$$G'_{i-1}(s) = (s - y_{i-1}) + \begin{cases} G'_i(s + \delta_{i-1}) & \text{If } s \leq s^*_i - \delta_{i-1} \\ 0 & \text{If } s^*_i - \delta_{i-1} < s \leq s^*_i \\ G'_i(s) & \text{If } s^*_i < s \end{cases} \tag{6}$$

The algorithm then finds s^*_i by finding zeros of G'_i. Starting from m, $G'_m = s - y_m$, and $s^*_m = y_m$. We design a special data structure, called *notable red-black trees*, for representing piecewise linear, continuous, strictly increasing functions. We initialize such a tree `T` to represent $G'_m(s) = s - y_m$. Assuming that at some time it represents G'_i, we need to support two operations:

1. Find the zero of G'_i to get s^*_i. Such an operation can be done efficiently $O(\log(m))$ time using a tree-like structure (Fig. 1 (a)).

2. Update `T` to represent G'_{i-1}. This operation is more complicated, but using the relation (6), we do the following: Split the interval containing s'_i. Move the left half of the piecewise linear function G'_i by δ_{i-1} (Fig. 1(b)), adding the constant zero function in between. Finally, we add the linear function $s - y_{i-1}$ to every interval, to get G'_{i-1}, which is again piecewise linear, continuous and strictly increasing.

To perform the operations in step (2) above, we cannot naïvely apply the transformations, `shift-by`(δ_{i-1}) and `add`$(s - y_{i-1})$ to every node in the tree, as it may take $O(m)$ operations. Instead, we simply leave a note (hence the name *notable* red-black trees) that such a transformation should be applied before the function is evaluated at that node or at any of its descendants. To prevent a large number of such notes accumulating at any given node we show that these notes satisfy certain commutative and additive relations, thus requiring us to keep track of no more than 2 notes at any given node. This lazy evaluation of notes allows us to perform all of the above operations in $O(\log(m))$ time. The details of the construction are provided in Appendix D.

5 Experiments

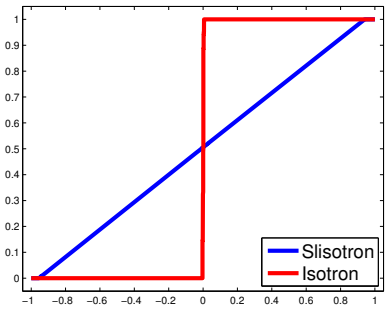
In this section, we present an empirical study of the SLISOTRON and GLM-TRON algorithms. We perform two evaluations using synthetic data. The first one compares SLISOTRON and Isotron [7] and illustrates the importance of imposing a Lipschitz constraint. The second one demonstrates the advantage of using SLISOTRON over standard regression techniques, in the sense that SLISOTRON can learn any monotonic Lipschitz function.

We also report results of an evaluation of SLISOTRON, GLM-TRON and several competing approaches on 5 UCI[12] datasets.

All errors are reported in terms of average root mean squared error (RMSE) using 10 fold cross validation along with the standard deviation.

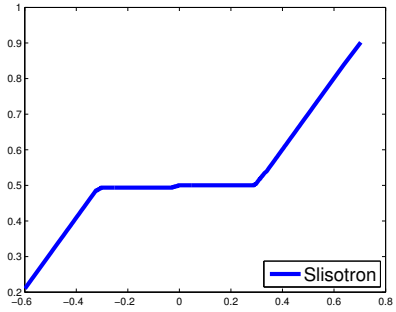
5.1 Synthetic Experiments

Although, the theoretical guarantees for Isotron are under the assumption that we get a fresh sample each round, one may still attempt to run Isotron on the same sample each iteration and evaluate the



SLISOTRON	Isotron	Δ
0.289 ± 0.014	0.334 ± 0.026	0.045 ± 0.018

(a) Synthetic Experiment 1



SLISOTRON	Logistic	Δ
0.058 ± 0.003	0.073 ± 0.006	0.015 ± 0.004

(b) Synthetic Experiment 2

Figure 2: (a) The figure shows the transfer functions as predicted by SLISOTRON and Isotron. The table shows the average RMSE using 10 fold cross validation. The Δ column shows the average difference between the RMSE values of the two algorithms across the folds. (b) The figure shows the transfer function as predicted by SLISOTRON. Table shows the average RMSE using 10 fold cross validation for SLISOTRON and Logistic Regression. The Δ column shows the average difference between the RMSE values of the two algorithms across folds.

empirical performance. Then, the main difference between SLISOTRON and Isotron is that while SLISOTRON fits the best *Lipschitz* monotonic function using LIR each iteration, Isotron merely finds the best monotonic fit using PAV. This difference is analogous to finding a large margin classifier vs. just a consistent one. We believe this difference will be particularly relevant when the data is sparse and lies in a high dimensional space.

Our first synthetic dataset is the following: The dataset is of size $m = 1500$ in $d = 500$ dimensions. The first co-ordinate of each point is chosen uniformly at random from $\{-1, 0, 1\}$. The remaining co-ordinates are all 0, except that for each data point one of the remaining co-ordinates is randomly set to 1. The true direction is $w = (1, 0, \ldots, 0)$ and the transfer function is $u(z) = (1 + z)/2$. Both SLISOTRON and Isotron put weight on the first co-ordinate (the true direction). However, Isotron overfits the data using the remaining (irrelevant) co-ordinates, which SLISOTRON is prevented from doing because of the Lipschitz constraint. Figure 2(a) shows the transfer functions as predicted by the two algorithms, and the table below the plot shows the average RMSE using 10 fold cross validation. The Δ column shows the average difference between the RMSE values of the two algorithms across the folds.

A principle advantage of SLISOTRON over standard regression techniques is that it is not necessary to know the transfer function in advance. The second synthetic experiment is designed as a sanity check to verify this claim. The dataset is of size $m = 1000$ in $d = 4$ dimensions. We chose a random direction as the "true" w and used a piecewise linear function as the "true" u. We then added random noise $(\sigma = 0.1)$ to the y values. We compared SLISOTRON to Logistic Regression on this dataset. SLISOTRON correctly recovers the true function (up to some scaling). Fig. 2(b) shows the actual transfer function as predicted by SLISOTRON, which is essentially the function we used. The table below the figure shows the performance comparison between SLISOTRON and logistic regression.

5.2 Real World Datasets

We now turn to describe the results of experiments performed on the following 5 UCI datasets: `communities`, `concrete`, `housing`, `parkinsons`, and `wine-quality`. We compared the performance of SLISOTRON (Sl-Iso) and GLM-TRON with logistic transfer function (GLM-t) against Isotron (Iso), as well as standard logistic regression (Log-R), linear regression (Lin-R) and a simple heuristic algorithm (SIM) for single index models, along the lines of standard iterative maximum-likelihood procedures for these types of problems (e.g., [13]). The SIM algorithm works by iteratively fixing the direction w and finding the best transfer function u, and then fixing u and

optimizing w via gradient descent. For each of the algorithms we performed 10-fold cross validation, using 1 fold each time as the test set, and we report averaged results across the folds.

Table 1 shows average RMSE values of all the algorithms across 10 folds. The first column shows the mean Y value (with standard deviation) of the dataset for comparison. Table 2 shows the average difference between RMSE values of SLISOTRON and the other algorithms across the folds. Negative values indicate that the algorithm performed better than SLISOTRON. The results suggest that the performance of SLISOTRON (and even Isotron) is comparable to other regression techniques and in many cases also slightly better. The performance of GLM-TRON is similar to standard implementations of logistic regression on these datasets. This suggests that these algorithms should work well in practice, while providing non-trivial theoretical guarantees.

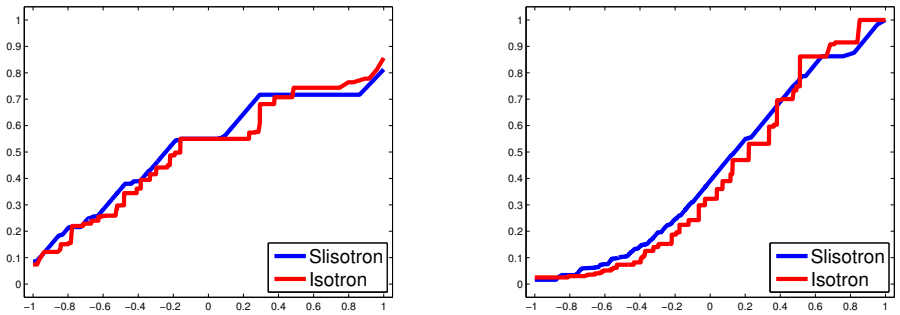
It is also illustrative to see how the transfer functions found by SLISOTRON and Isotron compare. In Figure 3, we plot the transfer functions for `concrete` and `communities`. We see that the fits found by SLISOTRON tend to be smoother because of the Lipschitz constraint. We also observe that `concrete` is the only dataset where SLISOTRON performs noticeably better than logistic regression, and the transfer function is indeed somewhat far from the logistic function.

Table 1: Average RMSE values using 10 fold cross validation. The $\bar{Y}$ column shows the mean Y value and standard deviation.

dataset	$\bar{Y}$	Sl-Iso	GLM-t	Iso	Lin-R	Log-R	SIM
communities	0.24 ± 0.23	0.13 ± 0.01	0.14 ± 0.01	0.14 ± 0.01	0.14 ± 0.01	0.14 ± 0.01	0.14 ± 0.01
concrete	35.8 ± 16.7	9.9 ± 0.9	10.5 ± 1.0	9.9 ± 0.8	10.4 ± 1.1	10.4 ± 1.0	9.9 ± 0.9
housing	22.5 ± 9.2	4.65 ± 1.00	4.85 ± 0.95	4.68 ± 0.98	4.81 ± 0.99	4.70 ± 0.98	4.63 ± 0.78
parkinsons	29 ± 10.7	10.1 ± 0.2	10.3 ± 0.2	10.1 ± 0.2	10.2 ± 0.2	10.2 ± 0.2	10.3 ± 0.2
winequality	5.9 ± 0.9	0.78 ± 0.04	0.79 ± 0.04	0.78 ± 0.04	0.75 ± 0.04	0.75 ± 0.04	0.78 ± 0.03

Table 2: Performance comparison of SLISOTRON with the other algorithms. The values reported are the average difference between RMSE values of the algorithm and SLISOTRON across the folds. Negative values indicate better performance than SLISOTRON.

dataset	GLM-t	Iso	Lin-R	Log-R	SIM
communities	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
concrete	0.56 ± 0.35	0.04 ± 0.17	0.52 ± 0.35	0.55 ± 0.32	-0.03 ± 0.26
housing	0.20 ± 0.48	0.03 ± 0.55	0.16 ± 0.49	0.05 ± 0.43	-0.02 ± 0.53
parkinsons	0.19 ± 0.09	0.01 ± 0.03	0.11 ± 0.07	0.09 ± 0.07	0.21 ± 0.20
winequality	0.01 ± 0.01	0.00 ± 0.00	-0.03 ± 0.02	-0.03 ± 0.02	0.01 ± 0.01



(a) `concrete` (b) `communities`

Figure 3: The transfer function u as predicted by SLISOTRON (blue) and Isotron (red) for the `concrete` and `communities` datasets. The domain of both functions was normalized to $[-1, 1]$.

References

[1] P. McCullagh and J. A. Nelder. *Generalized Linear Models (2nd ed.)*. Chapman and Hall, 1989.

[2] P. Hall W. Härdle and H. Ichimura. Optimal smoothing in single-index models. *Annals of Statistics*, 21(1):157–178, 1993.

[3] J. Horowitz and W. Härdle. Direct semiparametric estimation of single-index models with discrete covariates, 1994.

[4] A. Juditsky M. Hristache and V. Spokoiny. Direct estimation of the index coefficients in a single-index model. Technical Report 3433, INRIA, May 1998.

[5] P. Naik and C. Tsai. Isotonic single-index model for high-dimensional database marketing. *Computational Statistics and Data Analysis*, 47:775–790, 2004.

[6] P. Ravikumar, M. Wainwright, and B. Yu. Single index convex experts: Efficient estimation via adapted bregman losses. Snowbird Workshop, 2008.

[7] A. T. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT '09*, 2009.

[8] A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '05, pages 11–20, Washington, DC, USA, 2005. IEEE Computer Society.

[9] S. Shalev-Shwartz, O. Shamir, and K. Sridharan. Learning kernel-based halfspaces with the zero-one loss. In *COLT*, 2010.

[10] L. Yeganova and W. J. Wilbur. Isotonic regression under lipschitz constraint. *Journal of Optimization Theory and Applications*, 141(2):429–443, 2009.

[11] S. M. Kakade, A. T. Kalai, V. Kanade, and O. Shamir. Efficient learning of generalized linear and single index models with isotonic regression. arxiv.org/abs/1104.2018.

[12] UCI. University of california, irvine: `http://archive.ics.uci.edu/ml/`.

[13] S. Cosslett. Distribution-free maximum-likelihood estimator of the binary choice model. *Econometrica*, 51(3), May 1983.

[14] J. Shawe-Taylor and N. Christianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[15] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, 1999.

[16] T. Zhang. Covering number bounds for certain regularized function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

[17] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[18] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. In *NIPS*, 2010. (full version on arXiv).

[19] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

A Proof of Thm. 1

The reader is referred to GLM-TRON (Alg. 1) for notation used in this section.

The main lemma shows that as long as the error of the current hypothesis is large, the distance of our predicted direction vector w^t from the ideal direction w decreases by a substantial amount at each step.

Lemma 1. *At iteration t in* GLM-TRON*, suppose $\|w^t - w\| \leq W$, then if $\|(1/m)\sum_{i=1}^m (y_i - u(w \cdot x_i))x_i\|_w \leq \eta$, then*

$$\left\|w^t - w\right\|^2 - \left\|w^{t+1} - w\right\|^2 \geq \hat{\varepsilon}(h^t) - 5W\eta$$

Proof. We have

$$\left\|w^t - w\right\|^2 - \left\|w^{t+1} - w\right\|^2 \;=\; \frac{2}{m}\sum_{i=1}^m (y_i - u(w^t \cdot x_i))(w \cdot x_i - w^t \cdot x_i) - \left\|\frac{1}{m}\sum_{i=1}^m (y_i - u(w^t \cdot x_i))x_i\right\|^2. \tag{7}$$

Consider the first term above,

$$\frac{2}{m}\sum_{i=1}^m (y_i - u(w^t \cdot x_i))(w \cdot x_i - w^t \cdot x_i) \;=\; \frac{2}{m}\sum_{i=1}^m (u(w \cdot x_i) - u(w^t \cdot x_i))(w \cdot x_i - w^t \cdot x_i) + \frac{2}{m}\left(\sum_{i=1}^m (y_i - u(w \cdot x_i))x_i\right) \cdot (w - w^t).$$

Using the fact that u is non-decreasing and 1-Lipschitz (for the first term) and $\|w - w^t\| \leq W$ and $\|(1/m)\sum_{i=1}^m (y_i - u(w \cdot x_i))x_i\| \leq \eta$, we can lower bound this by

$$\frac{2}{m}\sum_{i=1}^m (u(w \cdot x_i) - u(w^t \cdot x_i))^2 - 2W\eta \;\geq\; 2\hat{\varepsilon}(h^t) - 2W\eta. \tag{8}$$

For the second term in (7), we have

$$\begin{aligned}
\left\|\frac{1}{m}\sum_{i=1}^m (y_i - u(w^t \cdot x_i))x_i\right\|^2 \;&=\; \left\|\frac{1}{m}\sum_{i=1}^m (y_i - u(w \cdot x_i) + u(w \cdot x_i) - u(w^t \cdot x_i))x_i\right\|^2 \\
\leq \left\|\frac{1}{m}\sum_{i=1}^m (y_i - u(w \cdot x_i))x_i\right\|^2 \;&+\; 2\left\|\frac{1}{m}\sum_{i=1}^m (y_i - u(w \cdot x_i))x_i\right\| \;\times\; \left\|\frac{1}{m}\sum_{i=1}^m (u(w \cdot x_i) - u(w^t \cdot x_i))x_i\right\| \\
&+ \left\|\frac{1}{m}\sum_{i=1}^m (u(w \cdot x_i) - u(w^t \cdot x_i))x_i\right\|^2
\end{aligned} \tag{9}$$

Using the fact that $\|(1/m)\sum_{i=1}^m (y_i - u(w \cdot x_i))x_i\| \leq \eta$, and using Jensen's inequality to show that $\|(1/m)\sum_{i=1}^m (u(w \cdot x_i) - u(w^t \cdot x_i))x_i\|^2 \leq (1/m)\sum_{i=1}^m (u(w \cdot x_i) - u(w^t \cdot x_i))^2 = \hat{\varepsilon}(h^t)$, and assuming $W \geq 1$, we get

$$\left\|\frac{1}{m}\sum_{i=1}^m (y_i - u(w \cdot x_i))x_i\right\|^2 \leq \hat{\varepsilon}(h^t) + 3W\eta \tag{10}$$

Combining (8) and (10) in (7), we get

$$\left\|w^t - w\right\|^2 - \left\|w^{t+1} - w\right\|^2 \geq \hat{\varepsilon}(h^t) - 5W\eta$$

□

To complete the proof of Theorem 1, the bound on $\hat{\varepsilon}(h^t)$ for some t now follows from Lemma 1. Let $\eta = 2(1 + \sqrt{log(1/\delta)/1})/\sqrt{m}$. Notice that $(y_i - u(w \cdot x_i))x_i$ for all i are i.i.d. 0-mean random variables with norm bounded by 1, so using Lemma 3 (Appendix B), $\|(1/m)\sum_{i=1}^m (y_i - u(w \cdot x_i))x_i\| \leq \eta$ with high probability. Now using Lemma 1, at each iteration of algorithm GLM-TRON, either $\left\|w^{t+1} - w\right\|^2 \leq \|w^t - w\|^2 - W\eta$, or $\hat{\varepsilon}(h^t) \leq 6W\eta$. If the

latter is the case, we are done. If not, since $\left\|w^{t+1}-w\right\|^2 \geq 0$, and $\left\|w^0-w\right\|^2 = \|w\|^2 \leq W^2$, there can be at most $W^2/(W\eta) = W/(\eta)$ iterations before $\hat{\varepsilon}(h^t) \leq 6W\eta$. Overall, there is some h^t such that

$$\hat{\varepsilon}(h^t) \leq O\left(\sqrt{\frac{W^2\log(1/\delta)}{m}}\right).$$

In addition, we can reduce this to a high-probability bound on $\varepsilon(h^t)$ using an uniform convergence argument (Lemma 5, Appendix B), which is applicable since $\|w^t\| \leq W$. Using a union bound, we get a bound which holds simultaneously for $\hat{\varepsilon}(h^t)$ and $\varepsilon(h^t)$.

B Proof of Thm. 2

This section is devoted to the proof of Thm. 2. The reader is referred to Section 4 for notation.

First we need a property of the LIR algorithm that is used to find the best one-dimensional non-decreasing 1-Lipschitz function. Formally, this problem can be defined as follows: Given as input $\langle\{z_i, y_i\}\rangle_{i=1}^m \in [-W, W] \times [0,1]$ the goal is to find $\hat{y}_1, \ldots, \hat{y}_m$ such that

$$\frac{1}{m}\sum_{i=1}^m (\hat{y}_i - y_i)^2, \tag{11}$$

is minimal, under the constraint that $\hat{y}_i = u(z_i)$ for some non-decreasing 1-Lipschitz function $u : [-W, W] \mapsto [0,1]$. After finding such values, LIR obtains an entire function u by interpolating linearly between the points. Assuming that z_i are in sorted order, this can be formulated as a quadratic problem with the following constraints:

$$\hat{y}_i - \hat{y}_{i+1} \leq 0 \qquad 1 \leq i < m \tag{12}$$

$$\hat{y}_{i+1} - \hat{y}_i - (z_{i+1} - z_i) \leq 0 \qquad 1 \leq i < m \tag{13}$$

Lemma 2. *Let $(z_1, y_1), \ldots, (z_m, y_m)$ be input to* LIR *where z_i are increasing and $y_i \in [0,1]$. Let $\hat{y}_1, \ldots, \hat{y}_m$ be the output of* LIR*. Let f be any function such that $f(\beta) - f(\alpha) \geq \beta - \alpha$, for $\beta \geq \alpha$, then*

$$\sum_{i=1}^m (y_i - \hat{y}_i)(f(\hat{y}_i) - z_i) \geq 0$$

Proof. We first note that $\sum_{j=1}^m (y_j - \hat{y}_j) = 0$, since otherwise we could have found other values for $\hat{y}_1, \ldots, \hat{y}_m$ which make (11) even smaller. So for notational convenience, let $\hat{y}_0 = 0$, and we may assume w.l.o.g. that $f(\hat{y}_0) = 0$. Define $\sigma_i = \sum_{j=i}^m (y_j - \hat{y}_j)$. Then we have

$$\sum_{i=1}^m (y_i - \hat{y}_i)(f(\hat{y}_i) - z_i) = \sum_{i=1}^m \sigma_i((f(\hat{y}_i) - z_i) - (f(\hat{y}_{i-1}) - z_{i-1})). \tag{14}$$

Suppose that $\sigma_i < 0$. Intuitively, this means that if we could have decreased all values $\hat{y}_{i+1}, \ldots \hat{y}_m$ by an infinitesimal constant, then the objective function (11) would have been reduced, contradicting the optimality of the values. This means that the constraint $\hat{y}_i - \hat{y}_{i+1} \leq 0$ must be tight, so we have $(f(\hat{y}_{i+1}) - z_{i+1}) - (f(\hat{y}_i) - z_i) = -z_{i+1} + z_i \leq 0$ (this argument is informal, but can be easily formalized using KKT conditions). Similarly, when $\sigma_i > 0$, then the constraint $\hat{y}_{i+1} - \hat{y}_i - (z_{i+1} - z_i) \leq 0$ must be tight, hence $f(\hat{y}_{i+1}) - f(\hat{y}_i) \geq \hat{y}_{i+1} - \hat{y}_i = (z_{i+1} - z_i) \geq 0$. So in either case, each summand in (14) must be non-negative, leading to the required result. □

We also use another result, for which we require a bit of additional notation. At each iteration of the SLISOTRON algorithm, we run the LIR procedure based on the training sample $(x_1, y_1), \ldots, (x_m, y_m)$ and the current direction w^t, and get a non-decreasing Lipschitz function u^t. Define

$$\forall i \;\; \hat{y}_i^t = u^t(w^t \cdot x_i).$$

Recall that w, u are such that $\mathbb{E}[y|x] = u(w \cdot x)$, and the input to the SLISOTRON algorithm is $(x_1, y_1), \ldots, (x_m, y_m)$. Define

$$\forall i \;\; \bar{y}_i = u(w \cdot x_i)$$

to be the expected value of each y_i. Clearly, we do not have access to $\bar{y}_i$. However, consider a hypothetical call to LIR with inputs $\langle (w^t \cdot x_i, \bar{y}_i) \rangle_{i=1}^m$, and suppose LIR returns the function $\tilde{u}^t$. In that case, define

$$\forall i \;\; \tilde{y}_i^t = \tilde{u}^t(w^t \cdot x_i).$$

for all i. Our proof uses the following proposition, which relates the values $\hat{y}_i^t$ (the values we can actually compute) and $\tilde{y}_i^t$ (the values we could compute if we had the conditional means of each y_i). The proof of Proposition 1 is somewhat lengthy and requires additional technical machinery, and is therefore relegated to Appendix C.

Proposition 1. *With probability at least $1-\delta$ over the sample $\{(x_i, y_i)\}_{i=1}^m$, it holds for any t that $\frac{1}{m}\sum_{i=1}^m |\hat{y}_i^t - \tilde{y}_i^t|$ is at most the minimum of*

$$\min\left\{ O\left(\left(\frac{dW^2 \log(Wm/\delta)}{m}\right)^{1/3}\right), O\left(\left(\frac{W^2 \log(m/\delta)}{m}\right)^{1/4}\right). \right\}$$

The third auxiliary result we need is the following, which is well-known (see for example [14], Section 4.1).

Lemma 3. *Suppose $z_1, \ldots, z_m$ are i.i.d. 0-mean random variables in a Hilbert space, such that $\Pr(\|x_i\| \leq 1) = 1$. Then with probability at least $1-\delta$,*

$$\left\| \frac{1}{m}\sum_{i=1}^m z_i \right\| \leq 2\left(\frac{1+\sqrt{\log(1/\delta)/2}}{\sqrt{m}}\right)$$

With these auxiliary results in hand, we can now turn to prove Thm. 2 itself. The heart of the proof is the following lemma, which shows that the squared distance $\|w^t - w\|^2$ between w^t and the true direction w decreases at each iteration at a rate which depends on the error of the hypothesis $\hat{\varepsilon}(h^t)$:

Lemma 4. *Suppose that $\|w^t - w\| \leq W$ and $\|(1/m)\sum_{i=1}^m (y_i - \bar{y}_i)x_i\| \leq \eta_1$ and $(1/m)\sum_{i=1}^m |\hat{y}_i^t - \tilde{y}_i^t| \leq \eta_2$. Then*

$$\|w^t - w\|^2 - \|w^{t+1} - w\|^2 \geq \hat{\varepsilon}(h^t) - 5W(\eta_1 + \eta_2)$$

Proof. We have

$$\begin{aligned}\|w^{t+1} - w\|_2^2 &= \|w^{t+1} - w^t + w^t - w\|_2^2 \\ &= \|w^{t+1} - w^t\|_2^2 + \|w^t - w\|_2^2 + 2(w^{t+1} - w^t)\cdot(w^t - w)\end{aligned}$$

Since $w^{t+1} - w^t = (1/m)\sum_{i=1}^m (y_i - \hat{y}_i^t)x_i$, substituting this above and rearranging the terms we get,

$$\|w^t - w\|^2 - \|w^{t+1} - w\|^2 = \frac{2}{m}\sum_{i=1}^m (y_i - \hat{y}_i^t)(w \cdot x_i - w^t \cdot x_i) - \left\| \frac{1}{m}\sum_{i=1}^m (y_i - \hat{y}_i^t)x_i \right\|^2. \quad (15)$$

Consider the first term above,

$$\frac{2}{m}\sum_{i=1}^m (y_i - \hat{y}_i^t)(w \cdot x_i - w^t \cdot x_i) = \left(\frac{2}{m}\sum_{i=1}^m (y_i - \bar{y}_i)x_i\right)\cdot(w - w^t) \quad (16)$$

$$+ \frac{2}{m}\sum_{i=1}^m (\bar{y}_i - \tilde{y}_i^t)(w \cdot x_i - w^t \cdot x_i) \quad (17)$$

$$+ \frac{2}{m}\sum_{i=1}^m (\tilde{y}_i^t - \hat{y}_i^t)(w \cdot x_i - w^t \cdot x_i) \quad (18)$$

The term (16) is at least $-2W\eta_1$, the term (18) is at least $-2W\eta_2$ (since $|(w - w^t)\cdot x_i| \leq W$). We thus consider the remaining term (17). Letting u be the true transfer function, suppose for a minute it is strictly increasing, so its inverse u^{-1} is well defined. Then we have

$$\frac{2}{m}\sum_{i=1}^m (\bar{y}_i - \tilde{y}_i^t)(w \cdot x_i - w^t \cdot x_i) = \frac{2}{m}\sum_{i=1}^m (\bar{y}_i - \tilde{y}_i^t)(w \cdot x_i - u^{-1}(\tilde{y}_i^t)) + \frac{2}{m}\sum_{i=1}^m (\bar{y}_i - \tilde{y}_i^t)(u^{-1}(\tilde{y}_i^t) - w^t \cdot x_i)$$

The second term in the expression above is positive by Lemma 2. As to the first term, it is equal to $\frac{2}{m}\sum_{i=1}^{m}(\bar{y}_i-\tilde{y}_i^t)(u^{-1}(\bar{y}_i)-u^{-1}(\tilde{y}_i^t))$, which by the Lipschitz property of u is at least $\frac{2}{m}\sum_{i=1}^{m}(\bar{y}_i-\tilde{y}_i^t)^2 = 2\hat{\varepsilon}(\tilde{h}^t)$. Plugging this in the above, we get

$$\frac{2}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i^t)(w\cdot x_i - w^t\cdot x_i) \geq 2\hat{\varepsilon}(\tilde{h}^t) - 2W(\eta_1+\eta_2) \tag{19}$$

This inequality was obtained under the assumption that u is strictly increasing, but it is not hard to verify that the same would hold even if u is only non-decreasing.

The second term in (15) can be bounded, using some tedious technical manipulations (see (9) and (10) in Appendix A), by

$$\left\|\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i^t)x_i\right\|^2 \leq \hat{\varepsilon}(h^t) + 3W\eta_1 \tag{20}$$

Combining (19) and (20)) in (15), we get

$$\left\|w^t-w\right\|^2 - \left\|w^{t+1}-w\right\|^2 \geq 2\hat{\varepsilon}(\tilde{h}^t) - \hat{\varepsilon}(h^t) - W(5\eta_1+2\eta_2) \tag{21}$$

Now, we claim that

$$\hat{\varepsilon}(\tilde{h}^t) - \hat{\varepsilon}(h^t) \geq -\frac{2}{m}\sum_{i=1}^{m}|\hat{y}_i^t-\tilde{y}_i^t| \geq -2\eta_2,$$

since

$$\begin{aligned}
\hat{\varepsilon}(\tilde{h}^t) &= \frac{1}{m}\sum_{i=1}^{m}(\tilde{y}_i^t-\bar{y}_i)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}(\tilde{y}_i^t-\hat{y}_i^t+\hat{y}_i^t-\bar{y}_i)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i^t-\bar{y}_i)^2 + \left(\frac{1}{m}\sum_{i=1}^{m}(\tilde{y}_i^t-\hat{y}_i^t)\right)(\tilde{y}_i^t+\hat{y}_i^t-2\bar{y}_i)
\end{aligned}$$

and we have that $|\tilde{y}_i^t+\hat{y}_i^t-2\bar{y}_i| \leq 2$. Plugging this into (21) leads to the desired result. □

The bound on $\hat{\varepsilon}(h^t)$ in Thm. 2 now follows from Lemma 4. Using the notation from Lemma 4, η_1 can be set to the bound in Lemma 3, since $\{(y_i-\bar{y}_i)x_i\}_{i=1}^m$ are i.i.d. 0-mean random variables with norm bounded by 1. Also, η_2 can be set to any of the bounds in Proposition 1. η_2 is clearly the dominant term. Thus, we get that Lemma 4 holds, so either $\left\|w^{t+1}-w\right\|^2 \leq \|w^t-w\|^2 - W(\eta_1+\eta_2)$, or $\hat{\varepsilon}(h^t) \leq 3W(\eta_1+\eta_2)$. If the latter is the case, we are done. If not, since $\left\|w^{t+1}-w\right\|^2 \geq 0$, and $\left\|w^0-w\right\|^2 = \|w\|^2 \leq W^2$, there can be at most $W^2/(W(\eta_1+\eta_2)) = W/(\eta_1+\eta_2)$ iterations before $\hat{\varepsilon}(h^t) \leq 6W\eta$. Plugging in the values for η_1,η_2 results in the bound on $\hat{\varepsilon}(h^t)$.

Finally, to get a bound on $\varepsilon(h^t)$, we utilize the following uniform convergence lemma:

Lemma 5. *Suppose that $\mathbb{E}[y|x] = u(\langle w,x\rangle)$ for some non-decreasing 1-Lipschitz u and w such that $\|w\| \leq W$. Then with probability at least $1-\delta$ over a sample $(x_1,y_1),\ldots,(x_m,y_m)$, the following holds simultaneously for any function $h(x) = \hat{u}(\hat{w}\cdot x)$ such that $\|\hat{w}\| \leq W$ and a non-decreasing and 1-Lipschitz function $\hat{u}$:*

$$|\varepsilon(h)-\hat{\varepsilon}(h)| \leq O\left(\sqrt{\frac{W^2\log(m/\delta)}{m}}\right).$$

The proof of the lemma uses a covering number argument, and is shown as part of the more general Lemma 7 in Appendix C. This lemma applies in particular to h^t. Combining this with the bound on $\hat{\varepsilon}(h^t)$, and using a union bound, we get the result on $\varepsilon(h^t)$ as well.

C Proof of Proposition 1

To prove the proposition, we actually prove a more general result. Define the function class

$$\mathcal{U} = \{u : [-W, W] \to [0, 1] : u \text{ 1-Lipschitz}\}.$$

and

$$\mathcal{W} = \{x \mapsto \langle x, w\rangle : w \in \mathbb{R}^d, \|w\| \le W\},$$

where d is possibly infinite (for instance, if we are using kernels).

It is easy to see that the proposition follows from the following uniform convergence guarantee:

Theorem 3. *With probability at least $1-\delta$, for any fixed $w \in \mathcal{W}$, if we let*

$$\hat{u} = \arg\min_{u\in\mathcal{U}} \frac{1}{m}\sum_{i=1}^{m}(u(w\cdot x_i) - y_i)^2,$$

and define

$$\tilde{u} = \arg\min_{u\in\mathcal{U}} \frac{1}{m}\sum_{i=1}^{m}(u(w\cdot x_i) - \mathbb{E}[y|x_i])^2,$$

then

$$\frac{1}{m}\sum_{i=1}^{m}|\hat{u}(w\cdot x_i) - \tilde{u}(w\cdot x_i)| \le O\left(\min\left\{\left(\frac{dW^{3/2}\log(Wm/\delta)}{m}\right)^{1/3} + \sqrt{\frac{W^2\log(m/\delta)}{m}}\ ,\ \left(\frac{W^2\log(m/\delta)}{m}\right)^{1/4}\right\}\right).$$

To prove the theorem, we use the concept of (∞-norm) covering numbers. Given a function class $\mathcal{F}$ on some domain and some $\epsilon > 0$, we define $\mathcal{N}_\infty(\epsilon, \mathcal{F})$ to be the smallest size of a *covering set* $\mathcal{F}' \subseteq \mathcal{F}$, such that for any $f \in \mathcal{F}$, there exists some $f' \in \mathcal{F}$ for which $\sup_x |f(x) - f'(x)| \le \epsilon$. In addition, we use a more refined notion of an ∞-norm covering number, which deals with an empirical sample of size m. Formally, define $\mathcal{N}_\infty(\epsilon, \mathcal{F}, m)$ to be the smallest integer n, such that for any $x_1, \ldots, x_m$, one can construct a covering set $\mathcal{F}' \subseteq \mathcal{F}$ of size at most n, such that for any $f \in \mathcal{F}$, there exists some $f' \in \mathcal{F}$ such that $\max_{i=1,\ldots,m} |f(x_i) - f'(x_i)| \le \epsilon$.

Lemma 6. *Assuming $m, 1/\epsilon, W \ge 1$, we have the following covering number bounds:*

1. $\mathcal{N}_\infty(\epsilon, \mathcal{U}) \le \frac{1}{\epsilon}2^{2W/\epsilon}$.
2. $\mathcal{N}_\infty(\epsilon, \mathcal{W}) \le \left(1 + \frac{2W}{\epsilon}\right)^d$.
3. $\mathcal{N}_\infty(\epsilon, \mathcal{U}\circ\mathcal{W}) \le \frac{2}{\epsilon}2^{4W/\epsilon}\left(1 + \frac{4W}{\epsilon}\right)^d$.
4. $\mathcal{N}_\infty(\epsilon, \mathcal{U}\circ\mathcal{W}, m) \le \frac{2}{\epsilon}(2m+1)^{1+8W^2/\epsilon^2}$

Proof. We start with the first bound. Discretize $[-W, W]\times[0,1]$ to a two-dimensional grid $\{-W + \epsilon a, \epsilon b\}_{a=0,\ldots,2W/\epsilon, b=0,\ldots,1/\epsilon}$. It is easily verified that for any function $u \in \mathcal{U}$, we can define a piecewise linear function u', which passes through points in the grid, and in between the points, is either constant or linear with slope 1, and $\sup_x |u(x) - u'(x)| \le \epsilon$. Moreover, all such functions are parameterized by their value at $-W$, and whether they are sloping up or constant at any grid interval afterward. Thus, their number can be coarsely upper bounded as $2^{2W/\epsilon}/\epsilon$.

The second bound in the lemma is a well known fact - see for instance pg. 63 in [15]).

The third bound in the lemma follows from combining the first two bounds, and using the Lipschitz property of u (we simply combine the two covers at an $\epsilon/2$ scale, which leads to a cover at scale ϵ for $\mathcal{U}\circ\mathcal{W}$).

To get the fourth bound, we note that by corollary 3 in [16]. $\mathcal{N}_\infty(\epsilon, \mathcal{W}, m) \le (2m+1)^{1+W^2/\epsilon^2}$. Note that unlike the second bound in the lemma, this bound is dimension-free, but has worse dependence on W and ϵ. Also, we have $\mathcal{N}_\infty(\epsilon, \mathcal{U}, m) \le \mathcal{N}_\infty(\epsilon, \mathcal{U}) \le \frac{1}{\epsilon}2^{2W/\epsilon}$ by definition of covering

numbers and the first bound in the lemma. Combining these two bounds, and using the Lipschitz property of u, we get

$$\frac{2}{\epsilon}(2m+1)^{1+4W^2/\epsilon^2}2^{4W/\epsilon}.$$

Upper bounding $2^{4W/\epsilon}$ by $(2m+1)^{4W^2/\epsilon^2}$, the the fourth bound in the lemma follows. □

Lemma 7. *With probability at least $1-\delta$ over a sample $(x_1,y_1),\ldots,(x_m,y_m)$ the following bounds hold simultaneously for any $w \in \mathcal{W}, u, u' \in \mathcal{U}$,*

$$\left|\frac{1}{m}\sum_{i=1}^{m}(u(w\cdot x_i)-y_i)^2 - \mathbb{E}\left[(u(w\cdot x)-y)^2\right]\right| \leq O\left(\sqrt{\frac{W^2\log(m/\delta)}{m}}\right),$$

$$\left|\frac{1}{m}\sum_{i=1}^{m}(u(w\cdot x_i)-\mathbb{E}[y|x_i])^2 - \mathbb{E}\left[(u(w\cdot x)-\mathbb{E}[y|x])^2\right]\right| \leq O\left(\sqrt{\frac{W^2\log(m/\delta)}{m}}\right),$$

$$\left|\frac{1}{m}\sum_{i=1}^{m}|u(w\cdot x_i)-u'(w\cdot x_i)| - \mathbb{E}\left[|u(w\cdot x)-u'(w\cdot x)|\right]\right| \leq O\left(\sqrt{\frac{W^2\log(m/\delta)}{m}}\right)$$

Proof. Lemma 6 tells us that $\mathcal{N}_\infty(\epsilon,\mathcal{U}\circ\mathcal{W},m) \leq \frac{2}{\epsilon}(2m+1)^{1+8W^2/\epsilon^2}$. It is easy to verify that the same covering number bound holds for the function classes $\{(x,y)\mapsto(u(w\cdot x)-y)^2 : u\in\mathcal{U}, w\in\mathcal{W}\}$ and $\{x\mapsto(u(w\cdot x)-\mathbb{E}[y|x])^2 : u\in\mathcal{U}, w\in\mathcal{W}\}$, by definition of the covering number and since the loss function is 1-Lipschitz. In a similar manner, one can show that the covering number of the function class $\{x\mapsto|u(w\cdot x)-u'(w\cdot x)| : u,u'\in\mathcal{U}, w\in\mathcal{W}\}$ is at most $\frac{4}{\epsilon}(2m+1)^{1+32W^2/\epsilon^2}$.

Now, one just need to use results from the literature which provides uniform convergence bounds given a covering number on the function class. In particular, combining a uniform convergence bound in terms of the Rademacher complexity of the function class (e.g. Theorem 8 in [17]), and a bound on the Rademacher complexity in terms of the covering number, using an entropy integral (e.g., Lemma A.3 in [18]), gives the desired result. □

Lemma 8. *With probability at least $1-\delta$ over a sample $(x_1,y_1),\ldots,(x_m,y_m)$, the following holds simultaneously for any $w\in\mathcal{W}$: if we let*

$$\hat{u}_w(\langle w,\cdot\rangle) = \arg\min_{u\in\mathcal{U}}\frac{1}{m}\sum_{i=1}^{m}(u(w\cdot x_i)-y_i)^2$$

denote the empirical risk minimizer with that fixed w, then

$$\mathbb{E}(\hat{u}_w(w\cdot x)-y)^2 - \inf_{u\in\mathcal{U}}\mathbb{E}(u(w\cdot x)-y)^2 \leq O\left(W\left(\frac{d\log(Wm/\delta)}{m}\right)^{2/3}\right),$$

Proof. For generic losses and function classes, standard bounds on the the excess error typically scale as $O(1/\sqrt{m})$. However, we can utilize the fact that we are dealing with the squared loss to get better rates. In particular, using Theorem 4.2 in [19], as well as the bound on $\mathcal{N}_\infty(\epsilon,\mathcal{U})$ from Lemma 6, we get that for any fixed w, with probability at least $1-\delta$,

$$\mathbb{E}(\hat{u}_w(w\cdot x)-y)^2 - \inf_{u\in\mathcal{U}}\mathbb{E}(u(w\cdot x)-y)^2 \leq O\left(W\left(\frac{\log(1/\delta)}{m}\right)^{2/3}\right).$$

To get a statement which holds simultaneously for any w, we apply a union bound over a covering set of $\mathcal{W}$. In particular, by Lemma 6, we know that we can cover $\mathcal{W}$ by a set $\mathcal{W}'$ of size at most $(1+2W/\epsilon)^d$, such that any element in $\mathcal{W}$ is at most ϵ-far (in an ∞-norm sense) from some $w'\in\mathcal{W}'$. So applying a union bound over $\mathcal{W}'$, we get that with probability at least $1-\delta$, it holds simultaneously for any $w'\in\mathcal{W}$ that

$$\mathbb{E}(\hat{u}_{w'}(\langle w',x\rangle)-y)^2 - \inf_{u}\mathbb{E}(u(\langle w',x\rangle)-y)^2 \leq O\left(W\left(\frac{\log(1/\delta)+d\log(1+2W/\epsilon)}{m}\right)^{2/3}\right). \tag{22}$$

Now, for any $w \in \mathcal{W}$, if we let w' denote the closest element in $\mathcal{W}'$, then $u(w \cdot x)$ and $u(\langle w', x\rangle)$ are ϵ-close *uniformly* for any $u \in \mathcal{U}$ and any x. From this, it is easy to see that we can extend (22) to hold for any $\mathcal{W}$, with an additional $O(\epsilon)$ element in the right hand side. In other words, with probability at least $1-\delta$, it holds simultaneously for any $w \in \mathcal{W}$ that

$$\mathbb{E}(\hat{u}_w(w \cdot x) - y)^2 - \inf_u \mathbb{E}(u(w \cdot x) - y)^2 \leq O\left(W\left(\frac{\log(2/\delta) + d\log(1+2W/\epsilon)}{m}\right)^{2/3}\right) + \epsilon.$$

Picking (say) $\epsilon = 1/m$ provides the required result. □

Lemma 9. *Let F be a convex class of functions, and let $f^* = \arg\min_{f \in F} \mathbb{E}[(f(x)-y)^2]$. Suppose that $\mathbb{E}[y|x] \in \mathcal{U} \circ \mathcal{W}$. Then for any $f \in F$, it holds that*

$$\mathbb{E}[(f(x)-y)^2] - \mathbb{E}[(f^*(x)-y)^2] \geq \mathbb{E}\left[(f(x)-f^*(x))^2\right] \geq \left(\mathbb{E}\left[|f(x)-f^*(x)|\right]\right)^2.$$

Proof. It is easily verified that

$$\mathbb{E}[(f(x)-y)^2] - \mathbb{E}[(f^*(x)-y)^2] = \mathbb{E}_x[(f(x)-\mathbb{E}[y|x])^2 - (f^*(x)-\mathbb{E}[y|x])^2]. \tag{23}$$

This implies that $f^* = \arg\min_{f \in F} \mathbb{E}[(f(x)-\mathbb{E}[y|x])^2]$.

Consider the L_2 Hilbert space of square-integrable functions, with respect to the measure induced by the distribution on x (i.e., the inner product is defined as $\langle f, f'\rangle = \mathbb{E}_x[f(x)f'(x)]$). Note that $\mathbb{E}[y|x] \in \mathcal{U} \circ \mathcal{W}$ is a member of that space. Viewing $\mathbb{E}[y|x]$ as a function $y(x)$, what we need to show is that

$$\|f-y\|^2 - \|f^*-y\|^2 \geq \|f-f^*\|^2.$$

By expanding, it can be verified that this is equivalent to showing

$$\langle f^*-y, f-f^*\rangle \geq 0.$$

To prove this, we start by noticing that according to (23), f^* minimizes $\|f-y\|^2$ over F. Therefore, for any $f \in F$ and any $\epsilon \in (0,1)$,

$$\|(1-\epsilon)f^* + \epsilon f - y\|^2 - \|f^*-y\|^2 \geq 0, \tag{24}$$

as $(1-\epsilon)f^* + \epsilon f \in F$ by convexity of F. However, the right hand side of (24) equals

$$\epsilon^2 \|f-f^*\|^2 + 2\epsilon\langle f^*-y, f-f^*\rangle,$$

so to ensure (24) is positive for any ϵ, we must have $\langle f^*-y, f-f^*\rangle \geq 0$. This gives us the required result, and establishes the first inequality in the lemma statement. The second inequality is just by convexity of the squared function. □

Proof of Thm. 3. We bound $\frac{1}{m}\sum_{i=1}^m |\hat{u}(w \cdot x) - \tilde{u}(w \cdot x)|$ in two different ways, one which is dimension-dependent and one which is dimension independent.

We begin with the dimension-dependent bound. For any fixed w, let u^* be $\arg\min_{u \in \mathcal{U}} \mathbb{E}(u(w \cdot x) - y)^2$. We have from Lemma 8 that with probability at least $1-\delta$, simultaneously for all $w \in \mathcal{W}$,

$$\mathbb{E}(\hat{u}(w \cdot x) - y)^2 - \mathbb{E}(u^*(w \cdot x) - y)^2 \leq O\left(W\left(\frac{d\log(Wm/\delta)}{m}\right)^{2/3}\right),$$

and by Lemma 9, this implies

$$\mathbb{E}[|\hat{u}(w \cdot x) - u^*(w \cdot x)|] \leq O\left(\left(\frac{dW^{3/2}\log(Wm/\delta)}{m}\right)^{1/3}\right). \tag{25}$$

Now, we note that since $u^* = \arg\min_{u \in \mathcal{U}} \mathbb{E}(u(w \cdot x) - y)^2$, then $u^* = \arg\min_{u \in \mathcal{U}} \mathbb{E}(u(w \cdot x) - \mathbb{E}[y|x])^2$ as well. Again applying Lemma 8 and Lemma 9 in a similar manner, but now with respect to $\tilde{u}$, we get that with probability at least $1-\delta$, simultaneously for all $w \in \mathcal{W}$,

$$\mathbb{E}[|\tilde{u}(w \cdot x) - u^*(w \cdot x)|] \leq O\left(\left(\frac{dW^{3/2}\log(Wm/\delta)}{m}\right)^{1/3}\right). \tag{26}$$

Combining (25) and (26), with a union bound, we have

$$\mathbb{E}[|\hat{u}(w \cdot x) - \tilde{u}(w \cdot x)|] \leq O\left(\left(\frac{dW^{3/2}\log(Wm/\delta)}{m}\right)^{1/3}\right).$$

Finally, we invoke the last inequality in Lemma 7, using a union bound, to get

$$\frac{1}{m}\sum_{i=1}^{m}|\hat{u}(w \cdot x) - \tilde{u}(w \cdot x)| \leq O\left(\left(\frac{dW^{3/2}\log(Wm/\delta)}{m}\right)^{1/3} + \sqrt{\frac{W^2\log(m/\delta)}{m}}\right).$$

We now turn to the dimension-independent bound. In this case, the covering number bounds are different, and we do not know how to prove an analogue to Lemma 8 (with rate faster than $O(1/\sqrt{m})$). This leads to a somewhat worse bound in terms of the dependence on m.

As before, for any fixed w, we let u^* be $\arg\min_{u\in\mathcal{U}}\mathbb{E}[(u(w \cdot x) - y)^2]$. Lemma 7 tells us that the empirical risk $\frac{1}{m}\sum_{i=1}^{m}(u(w \cdot x_i) - y_i)^2$ is concentrated around its expectation uniformly for any u, w. In particular,

$$\left|\frac{1}{m}\sum_{i=1}^{m}(\hat{u}(w \cdot x_i) - y_i)^2 - \mathbb{E}\left[\hat{u}(w \cdot x) - y)^2\right]\right| \leq O\left(\sqrt{\frac{W^2\log(m/\delta)}{m}}\right)$$

as well as

$$\left|\frac{1}{m}\sum_{i=1}^{m}(u^*(w \cdot x_i) - y_i)^2 - \mathbb{E}\left[(u^*(w \cdot x) - y)^2\right]\right| \leq O\left(\sqrt{\frac{W^2\log(m/\delta)}{m}}\right),$$

but since $\hat{u}$ was chosen to be the empirical risk minimizer, it follows that

$$\mathbb{E}\left[(\hat{u}(w \cdot x_i) - y_i)^2\right] - \mathbb{E}\left[(u^*(w \cdot x) - y)^2\right] \leq O\left(\sqrt{\frac{W^2\log(m/\delta)}{m}}\right),$$

so by Lemma 9,

$$\mathbb{E}\left[|u^*(w \cdot x) - \hat{u}(w \cdot x)|\right] \leq O\left(\left(\frac{W^2\log(m/\delta)}{m}\right)^{1/4}\right) \tag{27}$$

Now, it is not hard to see that if $u^* = \arg\min_{u\in\mathcal{U}}\mathbb{E}[(u(w\cdot x)-y)^2]$, then $u^* = \arg\min_{u\in\mathcal{U}}\mathbb{E}[(u(w\cdot x) - \mathbb{E}[y|x])^2]$ as well. Again invoking Lemma 7, and making similar arguments, it follows that

$$\mathbb{E}\left[|u^*(w \cdot x) - \tilde{u}(w \cdot x)|\right] \leq O\left(\left(\frac{W^2\log(m/\delta)}{m}\right)^{1/4}\right). \tag{28}$$

Combining (27) and (28), we get

$$\mathbb{E}\left[|\hat{u}(w \cdot x) - \tilde{u}(w \cdot x)|\right] \leq O\left(\left(\frac{W^2\log(m/\delta)}{m}\right)^{1/4}\right).$$

We now invoke Lemma 7 to get

$$\frac{1}{m}\sum_{i=1}^{m}|\hat{u}(w \cdot x_i) - \tilde{u}(w \cdot x_i)| \leq O\left(\left(\frac{W^2\log(m/\delta)}{m}\right)^{1/4}\right). \tag{29}$$

□

D Lipschitz Isotonic Regression

The reader is referred to Section 4.1 for notation used in this section. We first prove that the functions G_i defined in (Section 4.1 eq. (4)) are piecewise quadratic, differentiable everywhere and strongly convex.

Lemma 10. *For $i = 1, \ldots, m$, the functions G_i (eq. (Section 4.1 eq. (4)) are piecewise quadratic, differentiable everywhere and strictly convex.*

Proof. We prove this assertion by induction on $j = m - i$. This statement is true for $j = 0$ ($i = m$), since $G_m(s) = (1/2)(s - y_m)^2$. Recall that $\delta_i = z_{i+1} - z_i$ and $s_i^* = \min_s G_i(s)$.

Suppose that the assertion is true for i ($j = m - i$), then we shall prove the same for $i - 1$ ($j = m - i + 1$). Suppose $\hat{y}_{i-1}$ is fixed to s. Recall that $G_{i-1}(s)$ is the minimum possible error attained by fitting points $(z_{i-1}, y_{i-1}), (z_i, y_i), \ldots, (z_m, y_m)$ satisfying the Lipschitz and monotonic constraints and setting $\hat{y}_{i-1} = s$ (the prediction at the point z_{i-1}). For any assignment $\hat{y}_i = s'$ the least error is obtained by fitting the rest of the points is $G_i(s')$. Thus,

$$G_{i-1}(s) = \frac{1}{2}(s - y_m)^2 + \min_{s'} G_i(s')$$

subject to the conditions that $s' \geq s$ and $s' - s \leq \delta_{i-1}$.

Since G_i is piecewise quadratic, differentiable everywhere and strictly convex, G_i is minimum at s_i^* and is increasing on both sides of s_i^*. Thus, the quantity above is minimized by picking s' to be as close to s_i^* as possible without violating any constraint. Hence, we get the recurrence relation:

$$G_{i-1}(s) = \frac{1}{2}(s - y_{i-1})^2 + \begin{cases} G_i(s + \delta_{i-1}) & \text{If } s \leq s_i^* - \delta_{i-1} \\ G_i(s_i^*) & \text{If } s_i^* - \delta_{i-1} < s \leq s_i^* \\ G_i(s) & \text{If } s_i^* < s \end{cases} \tag{30}$$

By assumption for i, G_i is a piecewise quadratic function, say defined on intervals $(a_0 = -\infty, a_1], \ldots, (a_{k-1}, a_k], (a_k, a_{k+1}], \ldots, (a_{l-1}, a_l = \infty)$. Let $s_i^* \in (a_{k-1}, a_k]$ and note that $G_i'(s_i^*) = 0$. Define an intermediate function $F(s)$ which is defined on $l + 1$ intervals, $(-\infty, a_1 - \delta_{i-1}], (a_1 - \delta_{i-1}, a_2 - \delta_{i-1}], \ldots, (a_{k-1} - \delta_{i-1}, s_i^* - \delta_{i-1}], (s_i^* - \delta_{i-1}, s_i^*], (s_i^*, a_k], (a_k, a_{k+1}], \ldots, (a_{l-1}, \infty)$, where

1. $F(s) = G_i(s + \delta_{i-1})$ on intervals $(-\infty, a_1 - \delta_{i-1}], \ldots, (a_{k-1} - \delta_{i-1}, s_i^* - \delta_{i-1}]$
2. $F(s) = G_i(s_i^*)$ on interval $(s_i^* - \delta_{i-1}, s_i^*]$
3. $F(s) = G_i(s)$ on intervals $(s_i^*, a_k], \ldots, (a_{l-1}, \infty)$

Since G_i is piecewise quadratic, so is F on intervals as defined above (constant on the interval $(s_i^* - \delta_{i-1}, s_i^*]$). F is differentiable everywhere, since $G_i'(s_i^*) = 0$. Observe that $G_{i-1}(s) = (1/2)(s - y_{i-1})^2 + F(s)$, this makes $G_{i-1}(s)$ strictly convex, while retaining the fact that it is piecewise quadratic and differentiable everywhere. □

D.1 Algorithm

We now give an efficient algorithm for Lipschitz Isotonic Regression. As discussed already in Section 4.1, it suffices to find s_i^* for $i = 1, \ldots, m$. Once these values are known it is easy to find values $\hat{y}_1, \ldots \hat{y}_m$ that are optimal. The pseudocode for LIR algorithm is given in Alg. 3.

As discussed in Section 4.1 the algorithm maintains G_i', and finds s_i^* by finding the zero of G_i'. G_i' is piecewise linear, continuous and strictly increasing. We design a new data structure *notable red-black trees*[2] to store these piecewise linear, continuous, strictly increasing functions.

A notable red-black tree (henceforth NRB-Tree) stores a piecewise linear function. The disjoint intervals on which the function is defined have a natural order, thus the intervals serve as the *key* and the linear function is stored as the *value* at a node. Apart from storing these key-value pairs, the nodes of the tree also store some *notes* that are applied to the entire subtree rooted at that node, hence the name *notable* red-black trees. We assume that standard tree operations, i.e. finding and inserting nodes can be performed in $O(\log(m))$ time. There is a slight caveat that there may be *notes* at some of the nodes, but we show in Section D.2 how to *flush* all notes at a given node in constant

[2]Any tree-structure that uses only local rotations and guarantees $O(\log(m))$ worst-case or amortized *find* and *insert* operations may be used.

Algorithm 3 LIR: Lipschitz Isotonic Regression

Input: $(z_1, y_1), \ldots, (z_m, y_m)$

$s_i^* = 0, i = 1, \ldots m$
$\hat{y}_i = 0, i = 1, \ldots m$
NRB-Tree T$(s - y_m)$; // new T represents $s - y_m$.

$\delta_i = z_{i+1} - z_i, i = 1, \ldots, m$

for $i = m, \ldots, 1$ **do**
  $s_i^* =$ T.zero()
  if i $= 1$ **then**
    break;
  end if
  T.update$(s_i^*, \delta_{i-1}, y_{i-1})$
end for
$\hat{y}_i = s_1^*$
for $i = 2, \ldots, m$ **do**
  if $s_i^* > \hat{y}_{i-1} + \delta_{i-1}$ **then**
    $\hat{y}_i = \hat{y}_{i-1} + \delta_i$
  else if $s_i^* \geq \hat{y}_{i-1}$ **then**
    $\hat{y}_i = s_i^*$
  else
    $\hat{y}_i = \hat{y}_{i-1}$
  end if
end for

Output: $\hat{y}_1, \ldots, \hat{y}_m$

time. Thus, whenever a node is accessed for find or insert operations the notes are first applied to the node appropriately and then the standard tree operation can continue.

The algorithm initializes T to represent the linear function $s - y_m$. This can be performed very easily by just creating a root node, setting the interval (key) to be $(-\infty, \infty)$ and the linear function (value) to $s - y_m$, represented by coefficients $(1, -y_m)$. There are two function calls that algorithm LIR makes on T.

1. T.zero(): This finds the *zero* (the point where the function is 0) of the piecewise linear function G_i represented by the NRB-Tree T. Since the function is piecewise linear, continuous and strictly increasing, a unique zero exists and the tree structure is well-suited to find the zero in $O(\log m)$ time.
2. T.update$(s_i^*, \delta_{i-1}, y_i)$. Having obtained the zero of G_i, the algorithm needs to update T to now represent the G_{i-1}. This involves the following:
   (a) Suppose $(a_{k-1}, a_k]$ is the interval containing s_i^*. Split the interval $(a_{k-1}, a_k]$ into two intervals $(a_{k-1}, s_i^*]$ and $(s_i^*, a_k]$ (see Figure 1(a) in Section 4.1).
   This operation is easy to perform on a binary tree in $O(\log(m))$ time: It involves finding the node representing the interval $(a_{k-1}, a_k]$ that contains s_i^*, then modifying it to represent the interval $(a_{k-1}, s_i^*]$. Then a new node representing the interval $(s_i^*, a_k]$ (currently none of the intervals in the tree overlaps with this interval) is inserted and the linear function on this interval is set to be the same as the one on $(a_{k-1}, s_i^*]$.
   (b) All the intervals to the left of s_i^*, i.e. up to $(a_{k-1}, s_i^*]$ are to be moved to the left by δ_{i-1}. Formally, if the interval is $(a, a']$ and the linear function is (l, l') representing $ls + l'$, then after moving to the left by δ_{i-1}, the interval becomes $(a - \delta_{i-1}, a' - \delta_{i-1}]$ and the linear function becomes $(l, l' + l\delta_{i-1})$ (see Figure 1(b) in Section 4.1).
   The above operation may involve modifying up to $O(m)$ nodes in the tree. However rather than actually making the modifications we only leave a note at appropriate nodes that such a modification is to be performed. Whenever a note shift-by(δ_{i-1})

is left at a node, it is to be applied to the entire subtree. Thus, it suffices to leave only $O(\log(m))$ notes. The pseudo-code for this is provided in Section D.2.1. There is an additional caveat that a node where the note is left, may already have other notes, this issue is also discussed in Section D.2.

(c) A new interval $(s_i^* - \delta_{i-1}, s_i^*]$ is added to T to fill in the gap created by the shifting to the left in the above step (see Fig. 1(b) in Sec. 4.1). This operation is just addition of a new node and can be performed in $O(\log m)$ time, the linear function at this node is set to be 0.

(d) The function $s - y_{i-1}$ is added to be added to every node in the tree. This is again done by leaving a note $\texttt{add}(1, -y_{i-1})$ at the root node. Recall that a note at any node is applied to the entire subtree, thus in this case the note applies to every node in the tree. Thus this operation is performed in constant time.

At the end of the steps described above, the NRB-Tree T represents the function G'_{i-1}. The notes are flushed lazily as required whenever a node is accessed by find or insert operations.

D.2 Applying Notes

In this section we provide the method to apply notes to the NRB-Tree. As mentioned in the earlier section there are two kinds of notes that may be applied to any node: $\texttt{shift-by}(\delta)$ and $\texttt{add}(\alpha, \beta)$. Suppose a node has an interval $(a, a']$ and linear function $ls+l'$, then applying the notes $\texttt{shift-by}(\delta)$ and $\texttt{add}(\alpha, \beta)$, in that order, modifies the node so that the interval is now $(a - \delta, a' - \delta]$ and the linear function is $(l + \alpha)s + l\delta + l' + \beta$. Notice that applying the note $\texttt{add}(\alpha, \beta - \alpha\delta)$ followed by $\texttt{shift-by}(\delta)$ produces the same effect. Thus the two kinds of note commute in the following sense:

- $\texttt{shift-by}(\delta) \cdot \texttt{add}(\alpha, \beta) \iff \texttt{add}(\alpha, \beta - \delta\alpha) \cdot \texttt{shift-by}(\delta)$

It is also immediate that both kinds of notes are additive in the following sense:

- $\texttt{shift-by}(\delta) \cdot \texttt{shift-by}(\delta') \iff \texttt{shift-by}(\delta + \delta')$
- $\texttt{add}(\alpha, \beta) \cdot \texttt{add}(\alpha', \beta') \iff \texttt{add}(\alpha + \alpha', \beta + \beta')$

Thus, we can now ensure the following:

1. There are never more than 2 notes at any node, and there is at most one note of each kind.
2. If both kinds of notes are present, shift-by notes are always applied before add notes.

Whenever a new note is left at a node which has any existing notes, the number of notes is reduced to 2 and the correct order ensured by using the above commutative and additive relations. Since the number of notes is at most 2, this can be done in constant time.

Flushing Notes: Whenever a node is visited as part of the standard tree operations *find* and *insert*, the notes at any node that is visited are flushed. This involves modifying the key (interval) and value (linear function) to reflect the notes. The notes are then left at the children of this node (if there are any). Furthermore, if the children already have notes, the number of notes is reduced to 2 and proper order maintained. Since the number of children is 2, there is only a constant overhead. The tree operations can thus be performed in $O(\log(m))$ time despite the notes.

D.2.1 Applying $\texttt{shift-by}(\delta)$ notes

In step $\texttt{T.update}(s_i^*, \delta_{i-1}, y_i)$ there were two kinds of notes that were applied. The note $\texttt{add}(s, -y_{i-1})$ was easy since it was only applied to the root node. The $\texttt{shift-by}(\delta_{i-1})$ is slightly more involved. The path to the interval $(a_{k-1}, s_i^*]$ is traced and the nodes in the path which lie to the left of s_i^* are modified appropriately. A note is applied at the left child of such nodes. The path is of length at most $O(\log(m))$ and operation at any node takes constant time. The actual pseudocode is straightforward and is provided in Alg. 4.

Algorithm 4 Applying `shift-by`(δ)

$I = (a_{k-1}, s^*]$ // I is the interval
$n = r$ // Set n to be the root node
while $(n.\texttt{interval} \neq I)$ **do**
  if $(n.\texttt{interval} < I)$ **then**
    Modify the interval and function at n according to `shift-by`(δ)
    Apply note `shift-by`(δ) at $n.\texttt{left}$;
    $n = n.\texttt{right}$;
  else
    $n = n.\texttt{left}$;
  end if
end while

// A this step $n.\texttt{interval} = I$
Modify the interval and function at n according to `shift-by`(δ)