The Isotron Algorithm: High-Dimensional Isotonic Regression

Adam Tauman Kalai*
Microsoft Research
One Memorial Drive
Cambridge, MA

Ravi Sastry
College of Computing
Georgia Tech
Atlanta, GA

Abstract

The Perceptron algorithm elegantly solves binary classification problems that have a margin between positive and negative examples. Isotonic regression (fitting an arbitrary increasing function in one dimension) is also a natural problem with a simple solution. By combining the two, we get a new but very simple algorithm with strong guarantees. Our ISOTRON algorithm provably learns Single Index Models (SIM), a generalization of linear and logistic regression, generalized linear models, as well as binary classification by linear threshold functions. In particular, it provably learns SIMs with unknown mean functions that are nondecreasing and Lipschitz-continuous, thereby generalizing linear and logistic regression and linear-threshold functions (with a margin). Like the Perceptron, it is straightforward to implement and kernelize. Hence, the ISOTRON provides a very simple yet flexible and principled approach to regression.

1 Introduction

As a motivating example, imagine learning to predict whether a person has diabetes from n real-valued measurements, based on a batch of training data $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{R}^n \times \{0,1\}$. In binary classification, one would like an accurate predictor $h : \mathbb{R}^n \to \{0,1\}$ where $h(x) = 1$ indicates that that someone with attributes x is more likely to have diabetes. More useful would be the regression problem of predicting the probability $\Pr[y = 1|x]$ (more generally the *conditional mean function* $\mathbb{E}[y|x]$) based on their attributes x. We consider three problems of increasing difficulty:

1. **Perceptron problem**. The input is m labeled examples that are guaranteed to be separable by a linear threshold function (with a margin), and our goal is to find a (nearly) accurate linear separator.

2. **Idealized SIM problem**. The input is labeled examples $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{R}^n \times \mathbb{R}$ that are guaranteed to satisfy: $y_i = u(w \cdot x_i)$ for some $w \in \mathbb{R}^n$ and nondecreasing (Lipschitz continuous) $u : \mathbb{R} \to \mathbb{R}$. The goal is to find a (nearly) accurate such u, w.

3. **SIM problem**. The input is now independent examples $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{R}^n \times \mathbb{R}$ drawn independently from a distribution D, whose conditional mean function is of the form $\mathbb{E}_{(x,y)\sim D}[x|y] = u(w \cdot x)$ for some $w \in \mathbb{R}^n$ and nondecreasing (Lipschitz continuous) $u : \mathbb{R} \to \mathbb{R}$. The goal is to find a (nearly) accurate such u, w.

The most interesting problem for the Isotron is the SIM problem, as Perceptron already addresses the first, and the idealized SIM problem has fewer applications. However, we present the first two settings for clarity. Our algorithm and analysis are very much Perceptron-like. A key difference between the SIM problem and that of a *Generalized Linear Model* (GLM), is that in a GLM, u is known, e.g. $u = \frac{1}{1+e^{-x}}$ in case of logisitic regression.

We call w the *direction* and u the *mean function*. In all three problems, approximating the "true" direction (and mean function) is not possible in general – there may be multiple consistent such w and/or u. Instead, we focus on accuracy as measured by squared error (equivalent to classification error in the Perceptron problem).

We give a simple algorithm that is proven to solve the SIM problem *in polynomial time* analogous to how batch Perceptron algorithm [10] solves the Perceptron problem. Put another way, we learn SIMS in the probabilistic concept model of Kearns and Schapire [6]. Moreover, the algorithm is a simple combination of the Perceptron algorithm and Isotonic regression – its updates run in time $O(m \log m)$ instead of $O(m)$ for the Perceptron. It is easy to Kernelize and our bounds do not depend on the dimension of the space.

Related work. A large literature of related work exists for GLMs (see, e.g., [7]) which assume prior knowledge of u. For the SIM problem, there is also a body of work in Statistics (see, e.g. [4, 3]) whose aim is to identify the "correct" u, w. Several additional restrictions must be imposed on the model to ensure that this can be uniquely identified. (Following Kearns and Schapire's p-concept model [6], our goal is to find any u, w that accurately predict the true regression function in polynomial time.) Kalai [5] gives a polynomial-time algorithm for learning SIMs. However his (improper learner) outputs a branching program and the bounds depend heavily on the dimension of the problem. In

*Part of this research was done while the author was at the Georgia Institute of Technology, supported in part by NSF SES-0734780, an NSF CAREER award, and a SLOAN fellowship.

Machine Learning, the following approach is common: first a linear separator algorithm (e.g., SVM) is run to get a direction w, followed by a post-fitting of u using Isotonic regression or Platt calibration [8].

1.1 Formal results

For the analysis, all labeled examples will be assumed to lie in $(x, y) \in \mathbb{B}_n \times [0, 1]$, where $\mathbb{B}_n = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\}$ is the unit ball in n dimensions. In practice, the algorithm makes sense on any data set and is uniform scale-invariant.

In the **idealized SIM** problem, we assume that $y_i = u(w \cdot x_i)$ for some $w \in \mathbb{B}_n$ and $u : [-1, 1] \to [0, 1]$ that is nondecreasing and G-Lipschitz, i.e.,

$$0 \leq u(b) - u(a) \leq G(b - a) \text{ for all } 0 \leq a \leq b \leq 1.$$

For differentiable u, this is equivalent to $u'(z) \in [0, G]$ for $z \in [-1, 1]$. We will require that the mean functions be nondecreasing and G-Lipschitz for some $G \geq 0$.

In the **Perceptron problem**, we impose a *margin* assumption for a (linearly separable) data set. Let $\langle x_i, y_i \rangle_{i=1}^m \in \mathbb{B}_n \times \{0, 1\}$ be a data set. We say that the data has margin $\gamma > 0$ in direction $w \in \mathbb{B}_n$ if $w \cdot x_i \geq \gamma$ for each i with $y_i = 1$ and $w \cdot x_i \leq -\gamma$ for each i with $y_i = 0$. The Lipschitz-SIM is a natural generalization of the margin assumption, as depicted in Figure 1(a).

Observation 1 *The Perceptron problem with a γ-margin is a special case of the idealized SIM problem for a $G = (2\gamma)^{-1}$-Lipschitz continuous function.*

Proof: Take $u(z) = \begin{cases} 1 & z > \gamma \\ \frac{1}{2} + \frac{z}{2\gamma} & z \in [-\gamma, \gamma] \\ 0 & z < -\gamma \end{cases}$. ■

Error in these two problems is measured empirically. For $h : \mathbb{R}^n \to \mathbb{R}$, define,

$$\widehat{\text{err}}(h) = \frac{1}{m} \sum_{i=1}^{m} \big(h(x_i) - y_i\big)^2.$$

This matches the fraction of mistakes in the case where $y_i, h(x_i) \in \{0, 1\}$.

Our first theorem is about the empirical error of ISOTRON on idealized SIM problems.

Theorem 1 *Suppose $\langle x_i, y_i \rangle \in \mathbb{B}_n \times [0, 1]$ satisfy $y_i = u(w \cdot x_i)$ for monotonic G-Lipschitz u and $\|w\| \leq 1$. Then for $h^t(x) = u^t(w^t \cdot x)$ computed by* ISOTRON*,*

$$\sum_{t=1}^{\infty} \widehat{\text{err}}(h^t) \leq G^2.$$

In other words, the *total* of the errors after running for any number of rounds is at most G^2. Therefore, for any $\epsilon > 0$, there must be some $t \leq \lceil G^2/\epsilon \rceil$ that has $\widehat{\text{err}}(h^t) \leq \epsilon$. In practice, the algorithm will be executed for finitely many iterations and the h^t with minimal empirical error could be chosen. Our analysis follows the classic analysis of the Perceptron, which is completely analogous though much easier (see Theorem 3).

1.1.1 SIM theorem

Our main theorem is in fact for the SIM problem. In this setting, we have a distribution D over $\mathbb{B}_n \times [0, 1]$. The *conditional mean function*[1] is $f(x) = \mathbb{E}_{(x,y)\sim D}[y|x]$. We measure error of another classifier $h : \mathbb{B}_n \to \mathbb{R}$ in terms of expected squared error and ε-error:

$$\text{err}(h) = \mathbb{E}_{(x,y)\sim D}\big[(h(x) - y)^2\big]$$
$$\varepsilon(h) = \mathbb{E}_{(x,y)\sim D}\big[(f(x) - h(x))^2\big].$$

Note that expected squared error has a nice decomposition,

$$\text{err}(h) = \varepsilon(h) + \mathbb{E}_{(x,y)\sim D}\big[(f(x) - y)^2\big].$$

Also note that since the last term above does not depend on h, minimizing $\text{err}(h)$ and $\varepsilon(h)$ are equivalent.

Our main theorem is a statement that the class of of G-Lipschitz SIMs is efficiently learnable in the probabilistic concept model of Kearns and Schapire [6], which requires accurately learning the conditional mean function by a polynomial time function.

Theorem 2 *Suppose $\big\langle (x_i, y_i) \big\rangle_{i=1}^m \in \mathbb{B}_n \times [0, 1]$ satisfy $y_i = u(w \cdot x_i)$ for monotonic G-Lipschitz u and $\|w\| \leq 1$. There is a poly$(1/\epsilon, \log(1/\delta), n)$ time algorithm that, given any $\delta, \epsilon > 0$, with probability $\geq 1 - \delta$, it outputs $h(x) = \hat{u}(\hat{w} \cdot x)$ with*

$$\varepsilon(h) = \text{err}(h) - \text{err}(f) \leq \epsilon.$$

1.2 Algorithms

Consider first the case of $n = 1$ dimension and $w = 1$. In this case, a simple choice would be,

$$\text{PAV}\big((x_1, y_1), \ldots, (x_m, y_m)\big) = \operatorname*{arg\,min}_{\text{nondecreasing } u:\mathbb{R}\to\mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} \big(u(x_i) - y_i\big)^2.$$

This is essentially the problem of *Isotonic Regression* [9]. Let $\hat{y}_i = u(x_i)$. While the $\hat{y}_i$'s are uniquely determined, the rest of u is not uniquely determined. The Pool Adjacent Violator (PAV) algorithm computes such a u in time $O(m \log m)$. The algorithm sorts the data so that $x_1 \leq x_2 \ldots \leq x_m$ and then computes $\hat{y}_1 \leq \hat{y}_2 \ldots \leq \hat{y}_m$ to minimize the above displayed quantity, which can be done in $O(m)$ time. One simple property of Isotonic regression is the following calibration property.

Observation 2 *For any $z \in \mathbb{R}$, $\sum_{i:\hat{y}_i = z} (y_i - z) = 0$.*

The intuition behind this statement is simple. Consider the pool of examples that have $\hat{y}_i = z$. If z is not the average of the y_i's, then we could decrease the squared error by moving it some finite $\epsilon > 0$ towards the average, which we can do without violating monotonicity. With this calibration property in hand, it is relatively easy to derive the PAV algorithm. The data are partitioned into pools, depicted here by solid red lines, where the prediction assigned to each example is the average of the y's in its pool. Initially, each example is in its

[1]The notation $\mathbb{E}_{(x,y)\sim D}[y|x]$ can also be interpreted as $E[Y|X = x]$ for random variables (X, Y) jointly distributed according to D.

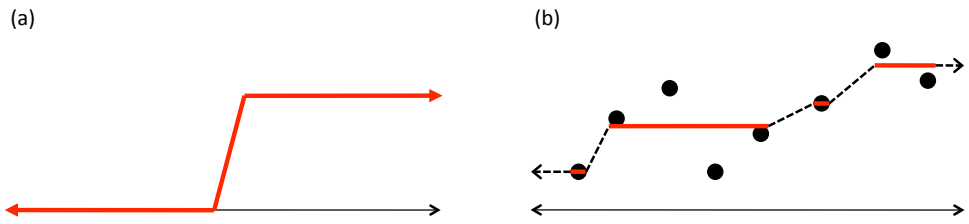

Figure 1: (a) The mean function corresponding to a linear separator with a margin. (b) An example of Isotonic regression.

own pool. Pools are then merged (in an arbitrary order) until the resulting function is nondecreasing. In between pools, one can perform linear interpolation. Figure 1(b) illustrates an example.

The rest of of u is chosen by, say, linear interpolation:

$$u(x) = \begin{cases} \hat{y}_1 & \text{if } x \le x_1 \\ \lambda \hat{y}_i + (1-\lambda)\hat{y}_{i+1} & \text{if } x = \lambda x_i + (1-\lambda)x_{i+1} \\ \hat{y}_m & \text{if } x \ge x_m \end{cases}.$$

The n-dimensional PERCEPTRON and ISOTRON algorithms are described below.

PERCEPTRON

Input: $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{R}^n \times \{0,1\}$
$w^1 := 0$
For $t := 1, 2, \ldots$:

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

where $u(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z \le 0 \end{cases}$.

ISOTRON

Input: $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{R}^n \times \mathbb{R}$
Let $w^1 := 0$
For $t := 1, 2, \ldots$:

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

where $u^t := \text{PAV}\big((x_1 \cdot w^t, y_1), \ldots, (x_m \cdot w^t, y_m)\big)$

Note that these algorithms are anytime algorithms – they are intended to be interrupted at any point at which point a classifier $h^t(x) = u^t(w^t \cdot x)$ may be output. Note also that for efficiency, the ISOTRON may perform interpolation only once at the end. For an actual implementation, one only maintains $\hat{y}_i^t = u^t(w^t \cdot x_i^t)$ on each iteration, which is all the PAV algorithm normally computes.

Kernelizing the ISOTRON is described in Section 3.

2 Analysis

We first briefly review the (batch) PERCEPTRON analysis.

Theorem 3 *Suppose $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{B}_n \times \{0,1\}$ is linearly separatable with margin $1/G$. Then for the h^t computed by the Perceptron algorithm, $\sum_{t=1}^\infty \widehat{\text{err}}(h^t) \le G^2$.*

The similarity to Theorem 1 should be clear. The elegant proof breaks into the following two elementary lemmas.

Lemma 1 *Suppose $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{B}_n \times \{0,1\}$ is linearly separated by w with margin $1/G$. Then for all $t \ge 1$, for the w^t, h^t computed by* PERCEPTRON ,

$$w^{t+1} \cdot w - w^t \cdot w \ge \widehat{\text{err}}(h^t)/G.$$

Proof: By definition of w^{t+1}, $(w^{t+1} - w^t) \cdot w = \frac{1}{m}\sum_{i=1}^m (y_i - \hat{y}_i^t)(x_i \cdot w)$. For each error on which $y_i \ne \hat{y}_i^t$, we have $(y_i - \hat{y}_i^t)(x_i \cdot w) \ge \frac{1}{G}$ because $|x_i \cdot w| \ge \frac{1}{G}$ and by assumption $y_i - \hat{y}_i^t = \text{sgn}(x_i \cdot w)$. This gives the lemma. ■

Lemma 2 *For all $t \ge 1$, for the w^t, h^t computed* PERCEPTRON ,

$$\|w^{t+1}\|^2 - \|w^t\|^2 \le \widehat{\text{err}}(h^t).$$

Proof: By definition of w^{t+1},

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

We next observe that $\sum_{i=1}^m (y_i - \hat{y}_i^t)x_i \cdot w^t \le 0$ because for each i on which $\hat{y}_i^t = 0$, $x_i \cdot w^t \le 0$, and for each i on which $\hat{y}_i^t = 1$, $x_i \cdot w^t > 0$. Finally, by the triangle inequality, we have,

$$\left\|\frac{1}{m}\sum (y_i - \hat{y}_i^t)x_i\right\|^2 \le \left(\frac{1}{m}\sum |y_i - \hat{y}_i^t|\,\|x_i\|\right)^2 = \left(\widehat{\text{err}}(h^t)\right)^2 \le \widehat{\text{err}}(h^t).$$

■

With these in hand, it is easy to prove Theorem 3.

Proof:[Theorem 3] By telescoping sums, Lemma 1 implies:

$$\sum_{t=1}^T \widehat{\text{err}}(h^t) \le G\left(w^{T+1} \cdot w\right) \le G\|w^{T+1}\|. \qquad (1)$$

Similarly, Lemma 2 implies $\|w^{T+1}\|^2 \le \sum_{t=1}^T \widehat{\text{err}}(h^t)$. Combining this with (1) gives,

$$\sum \widehat{\text{err}}(h^t) \le G\sqrt{\sum \widehat{\text{err}}(h^t)} \qquad (2)$$

This directly implies Theorem 3. ■

2.1 Idealized SIM analysis

In this section, we consider the simplified case where $y_i = f(x_i)$. While this case is of less practical interest and is easily solved by other means (finding such a consistent w can be formulated as a linear program), the analysis here conveys the main intuition for the full analysis but has fewer complications. The goal of this section is to prove Theorem 1. The proof is quite similar to the Perceptron analysis. Indeed, the statements of Lemmas 3 and 4 are nearly identical to Lemmas 1 and 2, but their proofs are significantly more involved. For ease of notation, we let $\hat{y}_i^t = u^t(w^t \cdot x_i)$ throughout the analysis.

Lemma 3 *Suppose $\langle (x_i, y_i) \rangle_{i=1}^m \in \mathbb{B}_n \times [0,1]$ satisfies $y_i = u(w \cdot x_i)$ for monotonic G-Lipschitz u and $\|w\| \leq 1$. Then for all $t \geq 1$, for the w^t, h^t computed by* ISOTRON*,*

$$w^{t+1} \cdot w - w^t \cdot w \geq \frac{1}{G} \widehat{\mathrm{err}}(h^t).$$

Proof:[Lemma 3] It will be helpful to consider the inverse of u, as seen later in eq. (4). Let $u([-1,1])$ be the range of u on inputs in $[-1,1]$, and let $v : u([-1,1]) \to [-1,1]$ be an inverse of u. Since there may be many inverses, for concreteness, we define:

$$v(y) = \inf\{z \in [-1,1] \mid u(z) = y\}.$$

By continuity of u, this exists and $u(v(y)) = y$ for all $y \in u([-1,1])$. Now, the remainder of the argument follows from the following (in)equalities, which are justified below:

$$\begin{aligned}
(w^{t+1} - w^t) \cdot w &= \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i^t)(x_i \cdot w) && (3)\\
&= \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i^t)(x_i \cdot w - v(\hat{y}_i^t)) && (4)\\
&\geq \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i^t) \frac{y_i - \hat{y}_i^t}{G} && (5)\\
&= \frac{\widehat{\mathrm{err}}(h^t)}{G} && (6)
\end{aligned}$$

Eq. (3) follows from the definition of w^{t+1}. For (4), we first need to verify that $v(\hat{y}_i^t)$ is well-defined. To see this, notice that $\hat{y}_i^t$ is always an average of some y_j's because of the calibration property of the PAV algorithm (see Observation 2). Hence $\hat{y}_i^t$ is in the set $u([-1,1])$, which is an interval. Second, we need to verify that the difference between (3) and (4) is 0, i.e.,

$$\sum_{i=1}^m (y_i - \hat{y}_i^t) v(\hat{y}_i^t) = 0.$$

To see this, we again use the calibration property of PAV. Consider the above sum over a single "pool" of examples. It must be that this sum is 0, $\sum (y_i - \hat{y}_i^t) v(\hat{y}_i^t) = 0$, because $v(\hat{y}_i^t)$ is constant across the pool and $\sum (y_i - \hat{y}_i^t) = 0$ by the calibration property. Hence, we have established (4). For (5), first consider the case that $y_i \geq \hat{y}_i^t$. Because u is nondecreasing and G-Lipschitz,

$$0 \leq y_i - \hat{y}_i^t = u(x_i \cdot w) - u\big(v(\hat{y}_i^t)\big) \leq G\big(x_i \cdot w - v(\hat{y}_i^t)\big).$$

Hence $(y_i - \hat{y}_i^t)(x_i \cdot w - v(\hat{y}_i^t)) \geq (y_i - \hat{y}_i^t)\frac{y_i - \hat{y}_i^t}{G}$. Similarly for the case of $y_i < \hat{y}_i^t$, hence (5). Finally (6) follows by definition of empirical error. ■

Lemma 4 *For all $t \geq 1$, for the w^t, h^t computed by* ISOTRON*,*

$$\|w^{t+1}\|^2 - \|w^t\|^2 \leq \widehat{\mathrm{err}}(h^t).$$

Proof:[Lemma 4] By definition of w^{t+1},

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

We next argue that

$$\sum_{i=1}^m (y_i - \hat{y}_i^t) x_i \cdot w^t \leq 0. \qquad (7)$$

To see this, we first claim that for any $\delta > 0$,

$$\sum_{i=1}^m \big(\hat{y}_i^t - y_i\big)^2 - \big(\hat{y}_i^t + \delta(x_i \cdot w^t) - y_i\big)^2 \leq 0.$$

This is true because $\hat{y}_i^t + \delta(x_i \cdot w^t)$ is also nondecreasing in $(x_i \cdot w^t)$ but $\hat{y}_i^t$ minimizes the sum of squared difference with respect to y_i over all such sequences. Rewriting this as a difference of squares gives,

$$\begin{aligned}
\sum \delta(x_i \cdot w^t)(2\hat{y}_i^t - 2y_i + \delta(x_i \cdot w^t)) \geq 0\\
\sum (x_i \cdot w^t)(\hat{y}_i^t - y_i + \frac{\delta}{2}(x_i \cdot w^t)) \geq 0
\end{aligned}$$

In the above, we have divided by $2\delta > 0$. But the above holds for every $\delta > 0$, hence it must hold for $\delta = 0$ by continuity, which is exactly (7).

Finally, by the triangle inequality

$$\begin{aligned}
\left\| \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i^t) x_i \right\|^2 &\leq \left(\frac{1}{m} \sum_{i=1}^m \left| y_i - \hat{y}_i^t \right| \|x_i\| \right)^2\\
&\leq \left(\frac{1}{m} \sum |y_i - \hat{y}_i^t| \right)^2.
\end{aligned}$$

By Holder's inequality, the last quantity is less than or equal to $\frac{1}{m} \sum (y_i - \hat{y}_i^t)^2 = \widehat{\mathrm{err}}(h^t)$. ■

Proof:[Theorem 1] Since Lemmas 3 and 4 match Lemmas 1 and 2, Theorem 1 follows exactly as Theorem 3 from equations (1) and (2). ■

2.2 General analysis (sketch)

The algorithm and analysis in this section are not meant to be optimal but rather to demonstrate that a variant of the ISOTRON algorithm, which we call ISOTRON II has theoretical guarantees for the SIM problem. We would expect that ISOTRON would work better in practice. Our modification uses fresh data in each iteration, hence it requires Tm examples.

ISOTRON II

Input: $T \geq 1$, $\left\langle (x_i^1, y_i^1) \right\rangle_{i=1}^m, \ldots, \left\langle (x_i^T, y_i^T) \right\rangle_{i=1}^m \in \mathbb{R}^n \times \mathbb{R}$
Let $w^1 := 0$
For $t := 1, 2, \ldots, T$:

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

where $u^t := \text{PAV}\big((x_1^t \cdot w^t, y_1^t), \ldots, (x_m^t \cdot w^t, y_m^t)\big)$

Recall that our goal is to find h with low $\varepsilon(h) = \text{err}(h) - \text{err}(f)$. As in the previous analysis, let $h^t(x) = u^t(w^t \cdot x)$ and let $\hat{y}_i^t = h^t(x_i^t)$. The following theorem says that, in expectation, the average ε over T iterations is low.

Theorem 4 *Let $G \geq 1, T \geq 1, m \geq (6T \log(eT)/G)^2$, and distribution D be over $\mathbb{B}_n \times [0,1]$ with conditional mean function $f(x) = u(w \cdot x)$ for nondecreasing G-Lipschitz $u : [-1,1] \to [0,1]$ and $w \in \mathbb{B}_n$. Then for h^t of the* ISOTRON II*,*

$$\mathbb{E}_{(x_1^1, y_1^1), \ldots (x_m^T, y_m^T) \sim D^{Tm}} \left[\sum_{t=1}^{T} \varepsilon(h^t) \right] \leq 8G^2.$$

Note that the above quantity is an expected err, i.e., an expectation over expectations. The proof of this theorem is rather involved and is in the appendix. The main idea is that the behavior of the algorithm will be statistically similar to as if it were in the idealized setting. This is combined with a generalization bound. In Theorem 2, we claimed a similar high-probability bound. The following standard trick can be used to convert low expected error to a high-probability bound.

Proof:[Theorem 2] We repeat the following $r = \lceil \lg(2/\delta) \rceil$ times. We run the Isotron II with $T \geq 16G^2/\epsilon$ and $m \geq (6T \log(eT)/G)^2$ on fresh data. Hence, the number of samples required is rTm. For each iteration, we take a random hypothesis h^t for t chosen uniformly random from $\{1, 2, \ldots, T\}$. This gives us a collection of r hypotheses, each with expected ε-error at most $\epsilon/4$. By Markov's inequality, with probability $1/2$, each one of these hypotheses has ε-error $\leq \epsilon/2$. Hence, with probability at most $2^{-r} \leq \delta/2$, none of the hypotheses have ε at most $\epsilon/2$. Otherwise, with probability $\geq 1 - \delta/2$, let us consider the case that at least one hypotheses has ε at most $\epsilon/2$.

Now, we draw a new set of $M = \log(2r/\delta)/\epsilon^2$ samples. Among these r hypotheses, we output one that has minimal empirical error on this new set. By our choice of M, by Chernoff-Hoeffding bounds, with probability $\leq \delta/2r$, each hypothesis has empirical error on this held-out set within $\epsilon/4$ of its true error. Assuming that this happens, we will therefore pick a hypothesis with $\varepsilon \leq \epsilon/2 + 2\epsilon/4 = \epsilon$. The total data requirements are $mrT + M$ and the algorithm runs in poly$(n, 1/\epsilon, \log(1/\delta))$ time. ■

3 Kernelization

The Kernelized version of the ISOTRON is quite simple and is given below. There is no regularization parameter. Instead, a held-out test set would be used in determining when to stop, to avoid overfitting.

Kernelized ISOTRON

Input: $\left\langle (x_i, y_i) \right\rangle_{i=1}^m \in X \times [0,1]$, Kernel $K : X \times X \to \mathbb{R}$
$\alpha^1 := (0, 0, \ldots, 0) \in \mathbb{R}^m$
For $t := 1, 2, \ldots$:
For $i = 1, 2, \ldots, m$:

$$\alpha_i^{t+1} := \alpha_i^t + y_i - u^t\left(z_i^t\right)$$

where $z_i^t = \sum_{j=1}^m \alpha_j K(x_i, x_j)$ and
$u^t = \text{PAV}\big((z_1^t, y_1), \ldots, (z_m^t, y_m)\big)$

In the algorithm above, the hypothesis on iteration t is $h^t(x) = u^t\left(\sum_{j=1}^m \alpha_j^t K(x, x_j)\right)$. Since there is no regularization, a held-out test set may be used to determine how many iterations to run, to avoid overfitting. Alternatively, we use the Kernelization approach of Blum, Balcan, and Vempala [2] which requires fewer support vectors. In this approach, we divide the training data into a set of B candidate support vectors $x_1, x_2, \ldots, x_B$, and the rest. We then treat the problem as a standard problem in $\mathbb{R}^B$ with a linear Kernel, where each example (training and test) is mapped to $\Phi(x) = \big(K(x, x_1), K(x, x_2), \ldots, K(x, x_B)\big)$. If one takes B to be significantly less than 1/2 of the training data, then we cannot overfit too much because we would have more than B examples for a model with B degrees of freedom.

4 Conclusions and future work

We have introduced a new method for learning SIMs that is simple and has appealing theoretical guarantees. The method inherits the properties of the Perceptron algorithm but is more general. From a theoretical point of view, it provides an interesting perspective on the properties of the Perceptron algorithm and Isotonic regression. Unfortunately, our analysis is batch, unlike the classic online analysis of the PERCEPTRON. It would be very interesting to be able to analyze an online variant of the ISOTRON. Also, thorough empirical work remains to compare the method to others in practice.

References

[1] Alon, N., and Spencer, J. (1992) The probabilistic method. New York: Wiley.

[2] Balcan, M., Blum, A., & Vempala, S. (2006). Kernels as Features: On Kernels, Margins, and Low-dimensional Mappings. *Machine Learning* 65, 79-94.

[3] Horowitz, J., & Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. em J. Amer. Statist. Assoc. 91, 1632-1640.

[4] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics* 58(1-2), 71-120.

[5] Kalai, A. (2004). Learning Monotonic Linear Functions. In *Proceedings of 17th Annual Conference on Learning Theory*, 487-501.

[6] Kearns, M., & Schapire, R. (1990). Efficient Distribution-free Learning of Probabilistic Concepts. *Journal of Computer and System Sciences* 48(3), 464-497.

[7] McCullagh, P., & Nelder, J. (1989) *Generalised Linear Models*. London: Chapman& Hall/CRC.

[8] Platt, J. (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classiffiers*, 61-74.

[9] Robertson, T., Wright, F., & Dykstra, R. (1988). *Order restricted statistical inference.* New York: John Wiley and Sons.

[10] Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review* 65(6), 386-408.

[11] Vapnik, V. & Chervonenkis, A. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.

A Proof of Theorem 4.

We now define an empirical version of ε,

$$\hat{\varepsilon}^t = \sum_{i=1}^{m} \left(f(x_i^t) - \hat{y}_i^t\right)^2.$$

We now give an analog of Lemma 3.

Lemma 5 *Let $G \geq 1, t \geq 1$, and distribution D be over $\mathbb{B}_n \times [0,1]$ with conditional mean function $f(x) = u(w \cdot x)$ for nondecreasing G-Lipschitz $u : [-1,1] \to [0,1]$ and $w \in \mathbb{B}_n$. Then for w^t of* ISOTRON II*,*

$$\mathbb{E}\left[w^{t+1} \cdot w - w^t \cdot w\right] \geq \frac{1}{G}\mathbb{E}[\hat{\varepsilon}^t] - 4\sqrt{\frac{2}{m}}.$$

In the above, expectations are over all data (x_j^j, y_i^j), for $1 \leq i \leq m, 1 \leq j \leq t$.

Proof: By definition of w^{t+1}, we have:

$$\begin{aligned}\left(w^{t+1} - w^t\right) \cdot w &= \frac{1}{m}\sum_{i=1}^{m}\left(y_i^t - \hat{y}_i^t\right) x_i^t \cdot w \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(f(x_i^t) - \hat{y}_i^t\right) x_i^t \cdot w + \\ &\qquad \left(y_i^t - f(x_i^t)\right) x_i^t \cdot w\end{aligned}$$

$$\mathbb{E}[\left(w^{t+1} - w^t\right) \cdot w] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i^t) - \hat{y}_i^t\right) x_i^t \cdot w\right] \quad (8)$$

For the last equality above, we have used the fact that $\mathbb{E}\left[\left(y_i^t - f(x_i^t)\right) x_i^t\right] = 0$, which follows from the definition of f. We now follow the approach of the proof of Lemma 3. In particular, we would like to consider the inverse of u on values $\hat{y}_i^t$. Note that $\hat{y}_i^t \in [0,1]$. There are two problems with this. First, the range of u is an interval $[a,b] \subseteq [0,1]$, but may not include $\hat{y}_i^t$. To address this, we define a new function, $U : [-2,2] \to [0,1]$. The three properties of U that we require are: (1) $U(t) = u(t)$ for all $t \in [-1,1]$, (2) U is G-Lipschitz and nondecreasing, and (3) $U(-2) = 0, U(2) = 1$, i.e., the range of U is the entire interval $[0,1]$. It is not hard to see that, for the domain $[-2,2]$, we have chosen, it is always possible to extend u (e.g., linearly) to such a function. The second problem, as in the proof of Lemma 3, is that the inverse of U (or u) is not necessarily unique at a point $z \in [0,1]$. As before, we consider the following inverse:

$$v : [0,1] \to [-2,2], v(z) = \inf\{x \in [-2,2] \mid U(x) = z\}.$$

Since U is continuous, we have that $U(v(z)) = z$ for all $z \in [0,1]$.

Now, by monotonicity and Lipschitz continuity,

$$\left(u(w \cdot x_i^t) - \hat{y}_i^t\right)\left(x_i^t \cdot w - v(\hat{y}_i^t)\right) \geq \frac{\left(u(w \cdot x_i^t) - \hat{y}_i^t\right)^2}{G}$$

$$\left(f(x_i^t) - \hat{y}_i^t\right) x_i^t \cdot w \geq \frac{\left(u(w \cdot x_i^t) - \hat{y}_i^t\right)^2}{G} + \left(u(w \cdot x_i^t) - \hat{y}_i^t\right) v(\hat{y}_i^t).$$

Taking expectations and combining with (8), gives

$$\mathbb{E}\left[\left(w^{t+1} - w^t\right) \cdot w\right] \geq \mathbb{E}\left[\frac{1}{Gm}\sum_{i=1}^{m}\left(f(x_i^t) - \hat{y}_i^t\right)^2\right] + \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(u(w \cdot x_i^t) - \hat{y}_i^t\right) v(\hat{y}_i^t)\right]$$

The first term in the RHS above is $\hat{\varepsilon}^t/G$. For the second term, note that the sequence $f(x_i^t) - y_i^t$ is an iid mean-0 sequence, while $v(\hat{y}_i^t)$ is a bounded nondecreasing sequence. Even if the adversary chose the latter after seeing the former, the two will probably not be very correlated. This is quantified precisely by Lemma 6, below, which implies a bound of,

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i^t) - \hat{y}_i^t\right) v(\hat{y}_i^t)\right] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i^t) - y_i^t\right) v(\hat{y}_i^t)\right] \geq -4\sqrt{\frac{2}{m}}.$$

■

Lemma 6 *For any integer $m \geq 1$ and reals $a < b$ let $X_1, X_2, \ldots, X_m$ be independent random variables, with $\mathbb{E}[X_i] = 0$, each in the bounded range $[-1,1]$. Let $A_1, \ldots, A_m$ be a sequence of random variables such that $a \leq A_1 \leq A_2 \leq \ldots \leq A_m \leq b$. Note that the A_i sequence need not be independent, and may depend on the X_i's as well. Then,*

$$\left|\mathbb{E}\left[\frac{A_1X_1 + \ldots + A_mX_m}{m}\right]\right| \leq (b-a)\sqrt{\frac{2}{m}}.$$

Proof: First consider the random variables $A_i' = \frac{A_i - a}{b-a} \in [0,1]$. Then, by linearity of expectation,

$$\mathbb{E}\left[\frac{A_1X_1 + \ldots + A_mX_m}{m}\right] = (b-a)\mathbb{E}\left[\frac{A_1'X_1 + \ldots + A_m'X_m}{m}\right] + a\mathbb{E}\left[\frac{X_1 + \ldots + X_m}{m}\right].$$

By the above and the fact that $\mathbb{E}\left[\frac{X_1+\ldots+X_m}{m}\right] = 0$, it suffices to prove the lemma for $a = 0, b = 1$. So WLOG $a = 0, b = 1$. We next claim that $\sum_{i=1}^m A_i X_i \leq \max_{0\leq j\leq m} \sum_{i=1}^j X_i$. To see this, note that:

$$\begin{aligned} A_1X_1 + \ldots + A_mX_m = \\ &\Delta_1(X_1 + X_2 + \ldots X_m)+ \\ &\Delta_2(X_2 + X_3 + \ldots + X_m) + \ldots \\ &+ \Delta_m X_m + \Delta_{m+1} 0 \end{aligned}$$

where $\Delta_1 = A_1, \Delta_2 = A_2 - A_1, \ldots, \Delta_m = A_m - A_{m-1}$, $\Delta_{m+1} = 1 - A_m$. Since $\Delta_i \geq 0, \sum_{i=1}^{m+1} \Delta_i = 1$, we have that $\sum_1^m A_i X_i$ is a convex combination of $\sum_1^j X_i$, over $j = 0, 1, \ldots, m$. Hence, it is no larger than the maximum. By Lemma 7 below, $\mathbb{E}[\max_{0\leq j\leq m} \sum_{i=1}^j X_i] \leq \sqrt{2m}$, we get the lemma. ■

Lemma 7 *Let $m \geq 1$ be any integer and $X_1, X_2, \ldots, X_n$ be independent random variables, with $\mathbb{E}[X_i] = 0$, each in the bounded range $[-1, 1]$. Then,*

$$\mathbb{E}\left[\max_{i\in\{0,1,\ldots,m\}} X_1 + X_2 + \ldots + X_i\right] \leq \sqrt{2m}.$$

Proof: Let Y be the random variable $Y = \max\{0, X_1, X_1 + X_2, \ldots, X_1+X_2+\ldots+X_m\}$. Next, we claim that, $\Pr[Y^2 \geq t] \leq e^{-t/(2m)}$, for all $t > 0$. Since, $Y \geq 0$, this is equivalent to $\Pr[Y \geq \sqrt{t}] \leq e^{-t/(2m)}$. To see this, fix t and consider the martingale $Z_1, Z_2, \ldots, Z_m$, where $Z_0 = 0$ and,

$$Z_i = \begin{cases} Z_{i-1} & \text{if } Z_{i-1} \geq \sqrt{t} \\ Z_{i-1} + X_i & \text{otherwise} \end{cases} \text{ for } i = 1, 2, \ldots, m-1.$$

Since $\mathbb{E}[X_i] = 0$, the Z_i's form a martingale. with bounded difference 1. Hence, Azuma's inequality [1] implies that $\Pr[Z_m \geq \sqrt{t}] \leq e^{-t/(2n)}$. Moreover, note that if $Y \geq \sqrt{t}$, then $Z_m \geq \sqrt{t}$. Hence, we have shown that $\Pr[Y^2 \geq t] \leq e^{-t/(2m)}$, for all $t > 0$. Now, for any nonnegative random variable $R \geq 0$, we have $\mathbb{E}[R] = \int_0^\infty \Pr[R \geq t]dt$. Hence,

$$\mathbb{E}[Y^2] = \int_0^\infty \Pr[Y^2 \geq t]dt \leq \int_0^\infty e^{-t/(2m)}dt = 2m.$$

Finally, using the fact that $\mathbb{E}[Y]^2 \leq \mathbb{E}\left[Y^2\right]$ for any random variable (non-negativity of variance), implies the lemma. ■

We now give the analog of Lemma 4.

Lemma 8 *Let $G \geq 1, t \geq 1$, and distribution D be over $\mathbb{B}_n \times [0, 1]$ with conditional mean function $f(x) = u(w \cdot x)$ for nondecreasing $u : [-1, 1] \to [0, 1]$ and $w \in \mathbb{B}_n$. Then for w^t of the* ISOTRON II*,*

$$\mathbb{E}\left[(w^{t+1})^2 - (w^t)^2\right] \leq \mathbb{E}\left[\hat{\varepsilon}^t\right] + \frac{3}{\sqrt{m}}.$$

In the above, expectations are over all data (x_j^j, y_i^j), for $1 \leq i \leq m, 1 \leq j \leq t$.

Proof: By expansion,

$$\begin{aligned}(w^{t+1})^2 =&(w^t)^2 + \frac{2}{m}\sum_{i=1}^m (y_i^t - \hat{y}_i^t)x_i^t \cdot w^t + \\ &\left(\frac{1}{m}\sum_{i=1}^m (y_i^t - \hat{y}_i^t)x_i^t\right)^2.\end{aligned}$$

Now, just as in the proof of Lemma 8, we have that,

$$\sum (y_i^t - \hat{y}_i^t)x_i^t \cdot w^t \leq 0.$$

Hence it remains to show that, $\left(\frac{1}{m}\sum_{i=1}^m (y_i^t - \hat{y}_i^t)x_i\right)^2 \leq 3/\sqrt{m}$. Next,

$$\left(\frac{1}{m}\sum_{i=1}^m (y_i^t - \hat{y}_i^t)x_i\right)^2 = (a + b)^2$$

where

$$\begin{aligned} a &= \frac{1}{m}\sum_{i=1}^m (f(x_i^t) - \hat{y}_i^t)x_i \\ b &= \frac{1}{m}\sum_{i=1}^m (y_i^t - f(x_i^t))x_i. \end{aligned}$$

By the triangle inequality,

$$\begin{aligned}\|a\|^2 &\leq \left(\frac{1}{m}\sum_{i=1}^m \left|f(x_i^t) - \hat{y}_i^t\right| \|x_i\|\right)^2 \\ &\leq \left(\frac{1}{m}\sum \left|f(x_i^t) - \hat{y}_i^t\right|\right)^2.\end{aligned}$$

By Holder's inequality, the last quantity is at most $\frac{1}{m}\sum(f(x_i^t) - \hat{y}_i^t)^2$, so $a^2 \leq \hat{\varepsilon}^t$. For b,

$$\begin{aligned} b^2 &= \frac{1}{m^2}\sum_{1\leq i,j\leq m} (y_i^t - f(x_i^t))(y_j^t - f(x_j^t))x_i^t \cdot x_j^t \\ &= \frac{1}{m^2}\sum_{i=1}^m (y_i^t - f(x_i^t))^2\|x_i^t\|^2. \end{aligned}$$

The cross terms above are 0 in expectation by independence. Since $(y_i^t - f(x_i^t))^2 \leq 1$ and $\|x_i^t\|^2 \leq 1$, we have that $\mathbb{E}[b^2] \leq 1/m$ and $\mathbb{E}[|b|] \leq 1/\sqrt{m}$. Finally, since $a^2 \leq \hat{\varepsilon}^t \leq 1$, we have,

$$\mathbb{E}\left[(a + b)^2\right] \leq \mathbb{E}\left[a^2 + 2|b| + b^2\right] \leq \mathbb{E}\left[\hat{\varepsilon}^t + \frac{2}{\sqrt{m}} + \frac{1}{m}\right]$$

This is at most $\mathbb{E}\left[\hat{\varepsilon}^t\right] + \frac{3}{\sqrt{m}}$.

■

What remains to prove Theorem 4 is now only a generalization bound for monotonic functions.

Lemma 9 *Let class C be the set of nondecreasing continuous functions $c : \mathbb{R} \to [0, 1]$. Let D be an arbitrary distribution on $(x, y) \in \mathbb{R} \times [0, 1]$ and $(x_1, y_1), \ldots, (x_m, y_m)$ be independent samples from D. Then*

$$\mathbb{E}_{(x_1,y_1),\ldots,(x_m,y_m)\sim D^m}[A] \leq 9\sqrt{\frac{\log(2m)}{m}}.$$

where

$$A = \sup_{c \in C} \left(\mathbb{E}_{(x,y)\sim D} \left[(c(x) - y)^2 \right] - \sum_{i=1}^{m} \frac{(c(x_i) - y_i)^2}{m} \right)$$

Proof: For any fixed $(x_1, y_1), \ldots, (x_m, y_m)$ we have,

$$\sup_{c \in C} \left(\mathbb{E}\left[(c(x) - y)^2 \right] - \sum_{i=1}^{m} \frac{(c(x_i) - y_i)^2}{m} \right) \leq \sup_{c \in C} \left(\mathbb{E}[c^2(x)] - \sum_{i=1}^{m} \frac{c^2(x_i)}{m} \right) + 2 \sup_{c \in C} \left(\sum_i \frac{c(x_i) y_i}{m} - \mathbb{E}[c(x)y] \right) + \left(\mathbb{E}[y^2] - \frac{1}{m} \sum_i y_i^2 \right).$$

In the above we have used the fact that $\sup_a F(a) + G(a) + z \leq \sup_a F(a) + \sup_a G(a) + z$, for any functions F, G and any $z \in \mathbb{R}$. Now, $\mathbb{E}[y^2 - \frac{1}{m}\sum_i y_i^2] = 0$ regardless of c, hence it suffices to show,

$$\mathbb{E}_{(x_i,y_i)} \left[\sup_{c \in C} \left(\mathbb{E}[c^2(x)] - \frac{1}{m} \sum_{i=1}^{m} c^2(x_i) \right) \right] \leq 3\sqrt{\frac{\log(2m)}{m}} \tag{9}$$

$$\mathbb{E}_{(x_i,y_i)} \left[\sup_{c \in C} \left(\frac{1}{m} \sum_i c(x_i) y_i - \mathbb{E}[c(x)y] \right) \right] \leq 3\sqrt{\frac{\log(2m)}{m}} \tag{10}$$

To this end, it will be helpful to consider the set of one-dimensional threshold functions,

$$\mathcal{I} = \{ f(x) = \mathbf{I}[x \geq \theta] \mid \theta \in \mathbb{R} \}.$$

Now, by standard VC-dimension bounds [11], since the VC-dimension of $\mathcal{I}$ is 1, we have that for any $\delta > 0$:

$$\Pr_{x_1,\ldots,x_m} \left[\sup_{\iota \in \mathcal{I}} \mathbb{E}_x[\iota(x)] - \sum_{i=1}^{m} \frac{\iota(x_i)}{m} \geq \sqrt{\frac{1 + \log \frac{8m}{\delta}}{m}} \right] \leq \delta.$$

Hence, for $\delta = m^{-1/2}$, we have that

$$\mathbb{E}_{x_1,\ldots,x_m} \left[\sup_{\iota \in \mathcal{I}} \mathbb{E}_x[\iota(x)] - \sum_{i=1}^{m} \frac{\iota(x_i)}{m} \right] \leq \delta + \sqrt{\frac{1 + \log \frac{8m}{\delta}}{m}} \leq 3\sqrt{\frac{\log(2m)}{m}}.$$

To establish (9), we use the fact that for any real $z \in [0, 1]$, $z = \int_0^1 \mathbf{I}[z \geq \theta] d\theta$. Thus,

$$\sup_{c \in C} \left(\mathbb{E}[c^2(x)] - \frac{1}{m} \sum_{i=1}^{m} c^2(x_i) \right) = \sup_{c \in C} \int_0^1 \mathbb{E}\big[\mathbf{I}[c^2(x) \geq \theta]\big] - \sum_{i=1}^{m} \frac{\mathbf{I}[c^2(x_i) \geq \theta]}{m} d\theta$$

$$\leq \int_0^1 \sup_{c \in C} \mathbb{E}\big[\mathbf{I}[c^2(x) \geq \theta]\big] - \sum_{i=1}^{m} \frac{\mathbf{I}[c^2(x_i) \geq \theta]}{m} d\theta$$

$$= \sup_{\iota \in \mathcal{I}} \mathbb{E}_x[\iota(x)] - \sum_{i=1}^{m} \frac{\iota(x_i)}{m} \tag{11}$$

$$\leq 3\sqrt{\frac{\log(2m)}{m}}$$

To see (11), note that $\mathbf{I}[c^2(\cdot) \geq \theta]$ is a function in $\mathcal{I}$, for any nondecreasing continuous c. This establishes (9).

For (10), we are going to consider a new distribution D' over $(x, y) \in \mathbb{R} \cup \{-\infty\} \times \{0, 1\}$. From D, we construct D' as follows. To get a sample from D', we take (x, y) from D and choose (x', y') as follows:

$$(x', y') = \begin{cases} (x, 1) & \text{with probability } y \\ (-\infty, 0) & \text{with probaility } (1 - y) \end{cases}.$$

It is not difficult to see that, for any $c \in C$, if we extend c to say that $c(-\infty) = 0$,

$$\mathbb{E}_{(x',y')\sim D'}[c(x')] = \mathbb{E}_{(x,y)\sim D}[c(x)y].$$

Next, we can imagine drawing a data set from D' by first drawing $(x_1, y_1), \ldots, (x_m, y_m)$ from D and then later the corresponding sample $(x'_1, y'_1), \ldots, (x_m, y'_m)$ from D'. For any $(x_1, y_1), \ldots, (x_m, y_m)$ and $c \in C$, we have

$$\mathbb{E}_{\langle (x'_i,y'_i) \rangle_{i=1}^m \sim D'^m} \left[\sum_{i=1}^{m} \frac{c(x'_i)}{m} \middle| \langle (x_i, y_i) \rangle_{i=1}^m \right] - \mathbb{E}_{(x',y')\sim D'}[c(x')] = \sum_{i=1}^{m} \frac{c(x_i)y_i}{m} - \mathbb{E}_{(x,y)\sim D}[c(x)y].$$

This holds for any c that is chosen based only on (x_i, y_i), independent of (x'_i, y'_i) since $\mathbb{E}[c(x'_i)|(x_i, y_i)] = c(x_i)y_i$. Hence, if we take supremums over $c \in C$ of both sides of the above displayed equation, the left can only be larger. It remains to show that,

$$\mathbb{E}_{(x'_1,y'_1),\ldots,(x'_m,y'_m)\sim D'^m}[A_{(x'_1,y'_1),\ldots,(x'_m,y'_m)}] \leq 3\sqrt{\frac{\log(2m)}{m}}.$$

where

$$A_{(x'_1,y'_1),\ldots,(x'_m,y'_m)} = \sup_{c \in C} \left(\frac{1}{m} \sum_i c(x'_i) - \mathbb{E}_{(x',y')\sim D}[c(x')] \right).$$

This follows in exactly the same manner as the chain of inequalities in (11). The main difference is that we need to consider intervals over $\mathbb{R} \cup \{-\infty\}$ such that $\iota(-\infty) = 0$. ■

The above lemma states that, for nondecreasing one-dimensional functions, the maximum expected difference between true and empirical errors is not large.

Proof:[Theorem 4] We first establish the following:

$$\mathbb{E}\left[\sum_{t=1}^{T} \hat{\varepsilon}^t \right] \leq 4G^2 \tag{12}$$

To do this, let $a = \mathbb{E}\left[\sum_{t=1}^{T} \hat{\varepsilon}^t\right]$. By our choice of m, $T/\sqrt{m} \leq G/6$. By Lemma 5, we have,

$$\mathbb{E}[w^{T+1} \cdot w] \geq \frac{1}{G}a - 4T\sqrt{\frac{2}{m}} \geq \frac{a}{G} - G.$$

By Lemma 8, we have,

$$\mathbb{E}\left[\|w^{T+1}\|\right]^2 \leq \mathbb{E}\left[\|w^{T+1}\|^2\right] \leq a + \frac{3T}{\sqrt{m}} \leq a + \frac{G}{2}.$$

Using these two facts in combination with the fact that $w^{T+1} \cdot w \leq \|w^{T+1}\| \, \|w\| \leq \|w^{T+1}\|$, we have:

$$\frac{a}{G} - G \leq \sqrt{a + \frac{G}{2}}.$$

If the left hand side above is negative, then we immediately have (12). Otherwise, squaring both sides and simplifying gives

$$\left(\frac{a}{G} - \frac{3G}{2}\right)^2 \leq \frac{G}{2} + \frac{9G^2}{4} \leq 3G^2.$$

This implies that $a/G \leq 3G/2 + G\sqrt{3}$ which impies (12).

This is what we need to establish (12). Finally, by Lemma 9, we have that, for each $t \leq T$, $\mathbb{E}[\varepsilon(h^t) - \hat{\varepsilon}^t] \leq 9\sqrt{\log(2m)/m}$. Therefore, we have,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{t} \varepsilon(h^t)\right] \leq \frac{4G^2}{T} + 9\sqrt{\frac{\log(2m)}{m}}. \qquad (13)$$

Now, $\log(2m)/m$ is decreasing in m for $m \geq 2$. (If $m \leq 2$, then $G^2/(36T^2) \geq 1/2$ and the theorem is trivial.) Otherwise, because $m \geq 9T^2 \log^2(eT)/G^2$, we have,

$$9\sqrt{\frac{\log(2m)}{m}} \leq \frac{3G}{2T\log(eT)}\sqrt{\log(72T^2\log^2(eT)/G^2)} \leq 4\frac{G}{T}.$$

In the above we have used the facts that $G \geq 1$ and

$$\sqrt{\log(72T^2\log^2(eT))} \leq 2.5\log(eT)$$

for all $T \geq 1$. The theorem follows from the above, eq. (13) and the fact that $G \geq 1$ and $G \leq G^2$. ■