Tossing Coins Under Monotonicity

Matey Neykov
Carnegie Mellon University
Department of Statistics & Data Science

Abstract

This paper considers the following problem: we are given n coin tosses of coins with monotone increasing probability of getting heads (success). We study the performance of the monotone constrained likelihood estimate, which is equivalent to the estimate produced by isotonic regression. We derive adaptive and non-adaptive bounds on the performance of the isotonic estimate, i.e., we demonstrate that for some probability vectors the isotonic estimate converges much faster than in general. As an application of this framework we propose a two step procedure for the binary monotone single index model, which consists of running LASSO and consequently running an isotonic regression. We provide thorough numerical studies in support of our claims.

1 INTRODUCTION

Recently there has been a lot of interest [see, e.g., Chatterjee et al., 2014, Bellec, 2018, Chatterjee et al., 2015, Guntuboyina and Sen, 2015, Zhang, 2002, among others] in the shape constrained Gaussian sequence model given by

$$Y_i = \theta_i + \varepsilon_i, \tag{1.1}$$

for $i \in [n] = \{1, \dots, n\}$ where the vector $\boldsymbol{\theta} \in \mathbb{R}^n$ is known to belong to a convex set K, and the noise follows a Gaussian distribution $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. A typical example of a convex set K is the set of monotone sequences, i.e., it is often assumed that the means satisfy $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_n$. Optimizing the least squares under the latter constraint is often called isotonic (or monotone) regression. Inspired by the recent work of Tian et al. [2017], in the present paper we consider a discretized version of this problem. Namely, we focus on the binary sequence model

$$O_i = Ber(p_i), i \in [n] \tag{1.2}$$

where we assume that $0 \leq p_1 \leq p_2 \leq \ldots \leq p_n \leq 1$.

Our goal is to derive results for model (1.2) that mirror those for the Gaussian model (1.1). To this end, we would like to explain the difference between the two models, and why it is non trivial to go from one to the other. Model (1.2) can be rewritten in the form

$$O_i = p_i + e_i,$$

where e_i is a mean 0, asymmetric (unless $p_i = \frac{1}{2}$) random variable with variance equal to $p_i(1 - p_i)$. Therefore the errors are independent but not identically distributed unlike in the Gaussian case. The majority of the recent results for the Gaussian sequence model rely heavily on two results from Bellec [2018], Chatterjee et al. [2014] – Theorems 2.3 and Theorem 1.1 respectively. Both of those results rely on the Gaussianity of the error terms. While Proposition 6.4 of Bellec [2018] relaxes the Gaussian assumption to a sub-Gaussian assumption, only symmetric noise is allowed which is not the case in model (1.2). In addition, Zhang [2002] has given conditions allowing for more general error terms (specifically for the isotonic regression case), but these conditions remain hard to verify in practice. To the best of our knowledge, although model (1.2) has been previously studied in the literature [Ayer et al., 1955, Banerjee and Wellner, 2001, Groeneboom and Wellner, 2012], rigorous results about adaptive and non-adaptive convergence rates of the isotonic estimate have not been established. In the present paper we extend results from the Gaussian case to the binary case, using classical symmetrization techniques.

As we mentioned, model (1.2) has been studied previously. It was motivated by different biological and statistical applications such as bio-assays [Ayer et al., 1955] and interval censoring [Banerjee and Wellner, 2001, Groeneboom and Wellner, 2012]. Here we will

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

motivate it with two additional perspectives – applications to propensity score estimation and binary choice models.

Suppose we observe the triple $(Y_i, X_i, A_i)_{i\in[n]}$ where Y_i indicates the outcome for the i^{th} patient, $X_i \in \mathbb{R}$ denotes a real valued covariate and $A_i \in \{0,1\}$ indicates whether the i^{th} patient was given a placebo or a new drug. In this setting $\mathbb{P}(A_i = 1|X_i)$ is the so called propensity score, which is of interest in a variety of causal inference applications, such as inverse probability weighting. Under the assumption that the likelihood of assigning the patient to the new drug increases as the covariate X_i increases, it is simple to see that propensity score estimation can be achieved by using model (1.2). This represents a non-parametric alternative to the more commonly used logistic regression. For more details on this application see Section 4.1 where we present numerical comparison between the two approaches.

In addition to the above, we apply our results of model (1.2) to study the following binary single index model (SIM)

$$Y_i = Ber(f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)), i \in [n], \tag{1.3}$$

where f is an unknown monotone increasing link function and $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is a high-dimensional s-sparse vector. Model (1.3) is also known as the binary choice model in the econometrics literature. Under the additional assumption that $\boldsymbol{X}_i$ follows a standard Gaussian distribution and f is Lipschitz, we propose a two step algorithm which is capable of estimating not only $\boldsymbol{\beta}^*$ (up to a proportionality constant) but also the function f. In the first step the procedure runs LASSO, and in the second step runs isotonic regression. This application has a similar flavor to SIM results with Gaussian designs such as those of [Plan and Vershynin, 2016, Plan et al., 2017, Neykov et al., 2016, Thrampoulidis et al., 2015] but unlike those works which focus solely on $\boldsymbol{\beta}^*$ estimation, we are also able to estimate f thanks to the isotonic regression step.

1.1 Summary of Results

In this section we informally state some selected results. The natural estimate for model (1.2) is the constrained likelihood estimator which solves the following program:

$$\widehat{\mathbf{p}} := \operatorname*{argmax}_{\mathbf{p}} \sum_{i\in[n]} O_i \log p_i + (1-O_i)\log(1-p_i)$$
$$\text{given that } 0 \le p_1 \le p_2 \le \ldots \le p_n \le 1.$$

We first note that this is equivalent to solving the isotonic regression

$$\operatorname*{argmin}_{\mathbf{p}} \sum_{i\in[n]} (O_i - p_i)^2 \text{ given that } p_1 \le p_2 \le \ldots \le p_n.$$

This is not a novel observation [see Part II Groeneboom and Wellner, 2012, e.g.], but we attach a short proof for completeness. We then study the estimation rates $\|\widehat{\mathbf{p}} - \mathbf{p}\|_2^2$. We show that depending on the underlying probability vector $\mathbf{p}$ there are two regimes of consistency for the isotonic estimate:

- In general
$$n^{-1}\|\widehat{\mathbf{p}} - \mathbf{p}\|_2^2 \lesssim n^{-2/3}$$
where $\lesssim$ denotes inequality up to constants.
- More specifically, for vectors $\mathbf{p}$ consisting of blocks of constant values, we show
$$n^{-1}\|\widehat{\mathbf{p}} - \mathbf{p}\|_2^2 \lesssim \frac{m(\mathbf{p})\log(en/m(\mathbf{p}))}{n},$$
where $m(\mathbf{p})$ denotes the number of unique values in the vector $\mathbf{p}$.

These two results parallel results established in the Gaussian sequence case [see Bellec, 2018, e.g.]. In addition to the above results, we derive a result regarding estimation rates of $\|\widehat{\mathbf{p}} - \mathbf{p}\|_p^p$ for any $1 \le p < 2$. We also show that those rates are minimax optimal (up to logarithmic factors). We then apply our theory to study model (1.3) where $\boldsymbol{X}_i$'s are assumed to have standard Gaussian design $\mathcal{N}(0, \mathbf{I})$ and f is assumed to be Lipschitz. We show that our two step procedure produces an estimate of $\boldsymbol{\beta}^*$: $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$, and an estimate of f: $\widehat{f}$ which satisfy

$$n^{-1}\sum_{i=1}^n (f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}))^2 \lesssim \frac{s\log p}{n} + \frac{1}{n^{2/3}}.$$

The latter inequality shows an interplay between two rates – the minimax rate of estimating $\boldsymbol{\beta}^*$ and the minimax rate of estimating a monotone function.

1.2 Notation

Here we outline some of the frequently used notation. Other notations will be defined as needed throughout the paper. For a sequence of numbers $X_1, \ldots X_n$ by $\overline{X}_{uv}$ we denote the average $\frac{\sum_{i=u}^v X_i}{v-u+1}$. For a vector $\mathbf{v} \in \mathbb{R}^p$ we use $\|\mathbf{v}\|_q$ to denote the ℓ_q norm (with the usual extension for $q = \infty$). For any integer $k \in \mathbb{N}$ we use the shorthand notation $[k] = \{1, \ldots, k\}$. We also use standard asymptotic notations. Given two sequences $\{a_n\}, \{b_n\}$ we write $a_n \lesssim b_n$ if there exists an absolute constant $C < \infty$ such that $a_n \le Cb_n$. We also sometimes write $a_n = O(b_n)$ if $a_n \lesssim b_n$.

1.3 Organization

The paper is structured as follows. In Section 2 we present our main findings regarding model (1.2). In Section 3 we apply the results of Section 2 to the binary SIM setting. In Section 4 we present thorough numerical studies. The discussion is deferred to the final Section 5.

2 TOSSING COINS UNDER MONOTONICITY

Suppose that we observe a single toss from each of n coins $Ber(p_i)$ for $i \in [n]$, where it is given that $0 \leq p_1 \leq \ldots \leq p_n \leq 1$. Let O_i be the result of the ith coin toss: 1 for heads (success) and 0 for tails (failure). Consider the following natural estimate of the vector $\mathbf{p}$:

$$\widehat{\mathbf{p}} := \underset{\mathbf{p}}{\operatorname{argmax}} \sum_{i\in[n]} O_i \log p_i + (1-O_i)\log(1-p_i) \quad \text{given } 0 \leq p_1 \leq p_2 \leq \ldots \leq p_n \leq 1. \tag{2.1}$$

Below we show that program (2.1) is in fact equivalent to optimizing monotone constrained least squares, which also known as isotonic regression.

Lemma 2.1 (Likelihood is Equivalent to Isotonic Regression)**.** The above estimate coincides with the estimate of

$$\widehat{\mathbf{p}} = \underset{\mathbf{p}}{\operatorname{argmin}} \sum_{i\in[n]} (O_i - p_i)^2 \text{ given } p_1 \leq p_2 \leq \ldots \leq p_n, \tag{2.2}$$

and therefore

$$\widehat{p}_i = \min_{v\geq i} \max_{u\leq i} \overline{O}_{uv} = \min_{v\geq i} \max_{u\leq i} \frac{\sum_{j=u}^{v} O_j}{v-u+1}. \tag{2.3}$$

Lemma 2.1 shows that program (2.1) is equivalent to running isotonic regression on O_i. Therefore (2.1) can be fitted very efficiently with the pool adjacent violators algorithm (PAVA) Mair et al. [2009]. We will now give a result which shows that when the vector $\mathbf{p}$ consists of blocks of constant values, then $\widehat{\mathbf{p}}$ adapts to the structure of $\mathbf{p}$ and attains nearly parametric rate of convergence. To this end, define the set of monotone sequences

$$\mathcal{S}_n^{\uparrow} = \{\mathbf{u} \in \mathbb{R}^n : u_1 \leq \ldots \leq u_n\}.$$

Theorem 2.2 (Misspecified Adaptive Rate)**.** We have that

$$\mathbb{E}\frac{\|\widehat{\mathbf{p}} - \mathbf{p}\|_2^2}{n} \leq \inf_{\mathbf{u}\in\mathcal{S}_n^{\uparrow}} \left[\frac{\|\mathbf{p}-\mathbf{u}\|_2^2}{n} + \frac{8\pi m(\mathbf{u}) \log(\frac{en}{m(\mathbf{u})})}{n}\right],$$

where $m(\mathbf{u})$ denotes the number of unique values in the vector $\mathbf{u}$.

In addition to showing the adaptivity of $\widehat{\mathbf{p}}$, Theorem 2.2 allows for model misspecification. The proof of Theorem 2.2 is presented in the end of this section. All other proofs are relegated to the supplement due to space considerations. Next we argue that $\widehat{\mathbf{p}}$ is consistent with high probability even when the vector $\mathbf{p}$ does not necessarily consist of blocks of constant values.

Theorem 2.3 (Misspecified Non-Adaptive Rate)**.** Let $\widehat{\mathbf{p}}$ be the estimate of (2.1). Then for any $\mathbf{u} \in \mathcal{S}_n^{\uparrow}$ we have

$$\frac{\|\widehat{\mathbf{p}} - \mathbf{p}\|_2^2}{n} \leq \frac{\|\mathbf{u}-\mathbf{p}\|_2^2}{n} + \frac{C(1+V(\mathbf{u}))^{2/3}}{n^{2/3}} + \frac{4x}{n},$$

with probability at least $1-e^{-x}$, where C is an absolute constant, and $V(\mathbf{u}) = \max_i u_i - \min_i u_i$.

Theorem 2.3 also allows for model misspecification. The difference between Theorems 2.2 and 2.3 is clear; when $\mathbf{p}$ has a simple structure with equal coefficients, the nearly parametric rates of Theorem 2.2 hold, and are much faster than the general rates attained by Theorem 2.3. Both Theorems 2.2 and 2.3 discuss the ℓ_2^2 loss. The following result discusses the ℓ_p^p losses for $1 \leq p < 2$ in the case when $\mathbf{p}$ consists of blocks of equal coefficients.

Theorem 2.4 (Well-Specified Adaptive ℓ_p^p Rates)**.** Fix $\mathbf{p} \in \mathcal{S}_n^{\uparrow}$ and a number $1 \leq p < 2$. Suppose that $\mathbf{p}$ consist of $m = m(\mathbf{p})$ blocks of equal coefficients – $N_1, \ldots, N_m$. Then

$$\mathbb{E}\frac{\|\widehat{\mathbf{p}} - \mathbf{p}\|_p^p}{n} \leq \frac{C_p}{n} \sum_{i=1}^{m} |N_m|^{1-p/2} \leq C_p \left(\frac{m}{n}\right)^{p/2},$$

for some constant C_p depending solely on p.

In contrast to Theorem 2.2, Theorem 2.4 does not have an extraneous logarithmic factor in the rate, which as we will see below is the minimax optimal rate for the ℓ_p^p loss. However, unlike Theorem 2.2, Theorem 2.4 does not allow for model misspecification.

2.1 Minimax optimality

Minimax optimality in the Gaussian case has been studied by Bellec and Tsybakov [2015] and Gao et al. [2017]. In this section we would like to verify that the same minimax bounds derived by Bellec and Tsybakov [2015] continue to hold for coin tossing, and therefore using isotonic regression is optimal (up to logarithmic factors). Moreover, it turns out that the rate proved in Theorem 2.4 is (constant) optimal for the ℓ_p^p loss in the range $1 \leq p < 2$.

Proposition 2.5 (Minimax Optimality)**.** Let $m < \frac{8\delta n}{1-2\delta} \wedge \sqrt[3]{n\frac{32(1-5/2\delta)^2}{\delta(1-2\delta)}}$ for some small fixed constant $\delta > 0$. Then for any $p \geq 1$ the minimax risk is at least

$$\inf_{\widehat{\mathbf{p}}} \sup_{\mathbf{p}\in\mathcal{S}^{\uparrow}_{n,\delta}\cap\mathcal{S}^{\uparrow}_{n}(m)} \mathbb{E}\frac{\|\widehat{\mathbf{p}}-\mathbf{p}\|_p^p}{n} \geq \frac{1}{32}\left[\frac{\delta(1-2\delta)m}{32n}\right]^{p/2},$$

where $\mathcal{S}^{\uparrow}_n(m) \subset \mathcal{S}^{\uparrow}_n$ is the set of all monotone sequences with m constant pieces, and $\mathcal{S}^{\uparrow}_{n,\delta} = \{\mathbf{u} \in \mathbb{R}^n : \delta \leq u_1 \leq \ldots \leq u_n \leq 1-\delta\}$.

2.2 Binomials

The goal of this section is to analyze what happens when instead of having individual Bernoulli observations O_i, we observe k coin tosses per each coin, i.e. we have

$$O_{ij} \sim Ber(p_i), j \in [k], i \in [n].$$

Do the estimation rates change when one uses

$$\overline{O}_i = k^{-1}\sum_{j\in[k]} O_{ij}$$

in the isotonic regression (2.2) as compared to using O_{ij}? We will show that indeed it is (slightly) better to use the average values $\overline{O}_i$ instead of simply using the original observations. First we will argue that while Theorem 2.2 remains valid it gives a slightly worse bound than the following proposition which uses the average number of successes $\overline{O}_i$ in place of the binary values O_{ij}.

Proposition 2.6 (Binomials Misspecified Adaptive Rate)**.** The estimates obtained from solving (2.2) with $\overline{O}_i$ instead of O_{ij} satisfy

$$\mathbb{E}\frac{\|\widehat{\mathbf{p}}-\mathbf{p}\|_2^2}{n} \leq \inf_{\mathbf{u}\in\mathcal{S}^{\uparrow}_n}\left[\frac{\|\mathbf{p}-\mathbf{u}\|_2^2}{n} + \frac{8\pi m(\mathbf{u})\log(\frac{en}{m(\mathbf{u})})}{kn}\right],$$

Remark 2.7. We remark that using the average probabilities $\overline{O}_i$ instead of using the $0,1$ representation O_{ij} improves the rate by a log factor. Assume that the vector $\widehat{\mathbf{p}}$ is obtained via averaging the estimated probabilities for each binomial after running an isotonic regression with O_{ij}. Note that by Jensen's inequality Theorem 2.2 guarantees that

$$\mathbb{E}\frac{\|\widehat{\mathbf{p}}-\mathbf{p}\|_2^2}{n} \leq \inf_{\mathbf{u}\in\mathcal{S}^{\uparrow}_n}\left[\frac{\|\mathbf{p}-\mathbf{u}\|_2^2}{n} + \frac{8\pi m(\mathbf{u})\log(\frac{ekn}{m(\mathbf{u})})}{kn}\right],$$

if we use the individual values O_{ij} in place of $\overline{O}_i$.

We also have an equivalent to Theorem 2.3.

Proposition 2.8 (Binomials Misspecified Non-Adaptive Rate)**.** For any $\mathbf{u} \in \mathcal{S}^{\uparrow}_n$ the estimates obtained from solving (2.2) with $\overline{O}_i$ instead of O_{ij} satisfy

$$\frac{\|\widehat{\mathbf{p}}-\mathbf{p}\|_2^2}{n} \leq \frac{\|\mathbf{u}-\mathbf{p}\|_2^2}{n} + \frac{C(1+V(\mathbf{u})\sqrt{k})^{2/3}}{kn^{2/3}} + \frac{4x}{kn},$$

with probability at least $1-e^{-x}$, where C is an absolute constant, and $V(\mathbf{u}) = \max_i u_i - \min_i u_i$.

Remark 2.9. Under the same assumptions as in Remark 2.7, Theorem 2.3 guarantees that with high probability

$$\frac{\|\widehat{\mathbf{p}}-\mathbf{p}\|_2^2}{n} \leq \frac{\|\mathbf{u}-\mathbf{p}\|_2^2}{n} + \frac{C(1+V(\mathbf{u}))^{2/3}}{(kn)^{2/3}} + \frac{4x}{kn},$$

if we use the individual values O_{ij} in place of $\overline{O}_i$. Hence the rate is improved when using the average values $\overline{O}_i$ for large values of k when $V(\mathbf{u})$ is small.

2.3 Proof of Theorem 2.2

Before we proceed with the proof we state a useful lemma.

Lemma 2.10 (Proposition 2.1 Bellec [2018])**.** Let $\mathcal{C}$ be a closed convex set in $\mathbb{R}^n$. Let $\boldsymbol{\theta} \in \mathbb{R}^n$ and suppose that $\boldsymbol{O} = \boldsymbol{\theta} + \mathbf{e}$ for some mean-zero random vector $\mathbf{e}$ such that $\mathbb{E}\|\mathbf{e}\|_2^2 < \infty$. Let $\widehat{\boldsymbol{\theta}} = \operatorname{argmin}_{\mathbf{v}\in\mathcal{C}}\|\boldsymbol{O}-\mathbf{v}\|_2^2$. Then

$$\mathbb{E}[\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_2^2] \leq \inf_{\mathbf{u}\in\mathcal{C}}\{\|\boldsymbol{\theta}-\mathbf{u}\|_2^2 + \mathbb{E}[\|\Pi_{\mathcal{T}_{\mathcal{C}}(\mathbf{u})}(\mathbf{e})\|_2^2]\},$$

where $\Pi_{\mathcal{T}_{\mathcal{C}}(\boldsymbol{\theta})}$ denotes the Euclidean projection of $\mathbf{e}$ onto the closed convex cone

$$\mathcal{T}_{\mathcal{C}}(\mathbf{u}) = \{t(\boldsymbol{\eta}-\mathbf{u}) : t \geq 0, \boldsymbol{\eta} \in \mathcal{C}\}.$$

Proof of Theorem 2.2. Lemma 2.10 shows that one needs to consider the projection on the cone $\mathcal{T}_{\mathcal{S}^{\uparrow}_n}(\mathbf{u})$. If $\mathbf{u}$ has $m = m(\mathbf{u})$ constant pieces – $N_1, \ldots, N_m$, Proposition 3.1 of Bellec [2018] shows that

$$\mathcal{T}_{\mathcal{S}^{\uparrow}_n}(\mathbf{u}) = \mathcal{S}^{\uparrow}_{|N_1|} \times \ldots \times \mathcal{S}^{\uparrow}_{|N_m|},$$

where $\mathcal{S}^{\uparrow}_l = \{\mathbf{t} \in \mathbb{R}^l : t_1 \leq \ldots \leq t_l\}$. By (B.10) of Amelunxen et al. [2014] we have that

$$\|\Pi_{\mathcal{T}_{\mathcal{C}}(\mathbf{u})}(\mathbf{e})\|_2^2 = \sum_{l\in[m]}\|\Pi_{\mathcal{S}^{\uparrow}_{|N_l|}}(\mathbf{e}_{N_l})\|_2^2,$$

where $\mathbf{e}_{N_l} = (O_i - p_i)_{i\in N_l}$ is the restriction of the vector $\mathbf{e} = (O_i - p_i)_{i\in[n]}$ to the set of coordinates N_l. Therefore the proof boils down to calculating the projection of $O_i - p_i$ for $i \in N_l$ onto the set $\mathcal{S}^{\uparrow}_{|N_l|}$. Let $\widetilde{O}_i$ be i.i.d. copies of O_i, independent of O_i, and ε_i be a sequence

of i.i.d. Rademacher random variables independent of all other randomness. Using symmetrization have

$$\begin{aligned}
\mathbb{E}\|\Pi_{\mathcal{S}^{\uparrow}_{|N_l|}}(\mathbf{e}_{N_l})\|_2^2 &= \mathbb{E}_{\mathbf{O}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}\sum_{i\in N_l}(O_i-p_i)t_i\Big]^2\\
&= \mathbb{E}_{\mathbf{O}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}\sum_{i\in N_l}(O_i-\mathbb{E}\widetilde{O}_i)t_i\Big]^2\\
&\le \mathbb{E}_{\mathbf{O},\widetilde{\mathbf{O}}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}\sum_{i\in N_l}(O_i-\widetilde{O}_i)t_i\Big]^2\\
&= \mathbb{E}_{\mathbf{O},\widetilde{\mathbf{O}},\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}\sum_{i\in N_l}\varepsilon_i(O_i-\widetilde{O}_i)t_i\Big]^2\\
&\le \frac{1}{2}\mathbb{E}_{\mathbf{O}}\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}2\sum_{i\in N_l}\varepsilon_i O_i t_i\Big]^2\\
&+ \frac{1}{2}\mathbb{E}_{\widetilde{\mathbf{O}}}\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}2\sum_{i\in N_l}\varepsilon_i \widetilde{O}_i t_i\Big]^2\\
&= \mathbb{E}_{\mathbf{O}}\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}2\sum_{i\in N_l}\varepsilon_i O_i t_i\Big]^2.
\end{aligned}$$

Next, for a vector of i.i.d. Gaussians $\xi_i, i\in[l]$ we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}2\sum_{i\in N_l}\varepsilon_i O_i t_i\Big]^2\\
&\le 4\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}2\sum_{i\in N_l}\varepsilon_i t_i\Big]^2\\
&= 8\pi\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}\sum_{i\in N_l}\varepsilon_i\mathbb{E}|\xi_i|t_i\Big]^2\\
&\le 8\pi\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\xi}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}\sum_{i\in N_l}\varepsilon_i|\xi_i|t_i\Big]^2\\
&= 8\pi\mathbb{E}_{\boldsymbol{\xi}}\Big[\sup_{\mathbf{t}\in\mathcal{S}^{\uparrow}_{|N_l|},\|\mathbf{t}\|_2\le 1}\sum_{i\in N_l}\xi_i t_i\Big]^2\\
&= 8\pi\sum_{i\in[|N_l|]}\frac{1}{i},
\end{aligned}$$

where the first inequality follows by the contraction principle [see Theorem 11.6 Boucheron et al., 2013, e.g.] and last equality follows by a well known fact for the monotone cone and Gaussian projections [Amelunxen et al., 2014, see (D.12)]. Since by Jensen's inequality

$$\mathbb{E}\|\Pi_{\mathcal{T}_{\mathcal{C}}(\mathbf{u})}(\mathbf{e})\|_2^2 \le 8\pi\sum_{l\in[m]}\sum_{i\in[|N_l|]}\frac{1}{i}\le 8\pi m\log(en/m),$$

the proof is complete. □

3 APPLICTION TO BINARY SINGLE INDEX MODELS

In this section we consider the following model

$$Y_i = Ber(f(\boldsymbol{X}_i^\top\boldsymbol{\beta}^*)), i\in[2n] \tag{3.1}$$

where $\boldsymbol{X}_i\sim\mathcal{N}(0,\mathbf{I})$, $\boldsymbol{\beta}^*\in\mathbb{R}^p$ is an s-sparse unit vector, i.e., $\|\boldsymbol{\beta}^*\|_2=1$, and $f:\mathbb{R}\mapsto[0,1]$ is an unknown, strictly monotone increasing and L-Lipschitz link function. Examples of f can be the logistic or probit link functions which are given by

$$f_{\text{logistic}}(x) := \frac{\exp(x)}{1+\exp(x)},\quad f_{\text{probit}}(x) := \Phi(x), \tag{3.2}$$

where Φ is the standard normal cdf. This model, also known as the binary choice model, is similar in spirit to 1-bit compressive sensing [Boufounos and Baraniuk, 2008], but the Bernoulli sampling introduces noise and the unknown function f further complicates the model. Our model assumptions and recent results [Plan and Vershynin, 2016, Plan et al., 2017, Neykov et al., 2016, Thrampoulidis et al., 2015] motivate running ℓ_1-regularized least squares (i.e., LASSO) to obtain a proportional estimate of $\boldsymbol{\beta}^*$. Before we proceed with formalizing our two step procedure (which also aims at estimating f) we will first briefly sketch why one would expect to obtain a proportional estimate after running the least squares. To see why, it is convenient to consider the population version of the problem. We have the following simple result

Lemma 3.1 (Least Squares Proportionality)**.** Suppose we are given samples from model (3.1) with a strictly monotone increasing f. Then we have

$$\operatorname*{argmin}_{\boldsymbol{\beta}}\mathbb{E}(Y-\boldsymbol{X}^\top\boldsymbol{\beta})^2 = c_0\boldsymbol{\beta}^*, \tag{3.3}$$

where $c_0 := \mathbb{E}Y\boldsymbol{X}^\top\boldsymbol{\beta}^* = \mathbb{E}f(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{\beta}^* > 0$.

After obtaining a proportional estimate of $\boldsymbol{\beta}^*$ – $\widehat{\boldsymbol{\beta}}$, we can approximately sort the observations according to the monotonicity of $\boldsymbol{X}_i^\top\widehat{\boldsymbol{\beta}}$, since the function f is monotone increasing. In view of our results from the previous sections, this suggest running an isotonic regression on the sorted values. We are now ready to state our two step algorithm. The goal of the remainder of the section is to prove that the following procedure works well to recover both $\boldsymbol{\beta}^*$ and f.

- Split the data. Run LASSO on the first half:

$$\widetilde{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}}\sum_{i=1}^{n}(Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta})^2+\lambda\|\boldsymbol{\beta}\|_1,$$

  and normalize

$$\widehat{\boldsymbol{\beta}} := \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|_2}.$$

- Sort the second half of the data. Specifically, let π be the permutation of $\{n+1,\dots,2n\}$ so that $\boldsymbol{X}_{\pi_i}^\top\widehat{\boldsymbol{\beta}}\le\boldsymbol{X}_{\pi_{i+1}}^\top\widehat{\boldsymbol{\beta}}$ for $i\in[n]$ (breaking ties arbitrarily). Fit isotonic regression using $\boldsymbol{Y}_\pi =$

$(Y_{\pi_1}, \ldots, Y_{\pi_n})^\top$:

$$\widehat{\mathbf{f}}_\pi := \operatorname*{argmin}_{\mathbf{f} \in \mathcal{S}_n^\uparrow} \|\mathbf{Y}_\pi - \mathbf{f}\|_2^2.$$

Our goal is to show that the predicted estimate $\widehat{f}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})$ is close to $f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)$ for $i \in \{n+1, \ldots, 2n\}$. To compare how close is $\widehat{f}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})$ to $f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*)$ we will use the "in-sample" square loss function $\frac{1}{n}\sum_{i=n+1}^{2n}(f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}))^2$. Define the shorthand notation:

$$n_{p,s} := \frac{n}{s \log p}.$$

We have the following result, which shows that the in-sample square loss is small provided that $n_{p,s}$ is sufficiently large, and the tuning parameter λ is set appropriately.

Theorem 3.2 (Binary SIM Two-Step Procedure Guarantee)**.** Suppose that $n_{p,s}^{-1} = o(1)$,

$$n_{p,s} \gtrsim \frac{\mathbb{E}(Y - c_0 \boldsymbol{X}^\top \boldsymbol{\beta}^*)^2}{\lambda^2 s}, \tag{3.4}$$

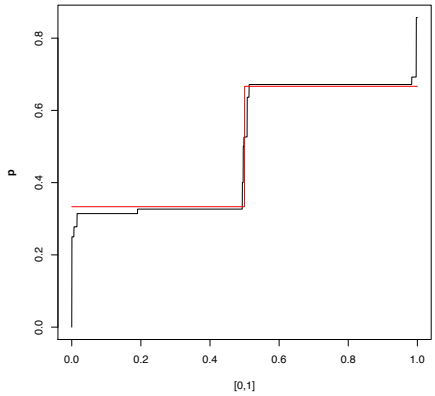
and that $\mathbb{E}Y^4 < \infty$ is fixed and does not scale with n. Then, assuming that $s \to \infty$ as $n \to \infty$ with overwhelming probability (i.e., at least .99) we have

$$\frac{1}{n}\sum_{i=n+1}^{2n}(f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}))^2 \lesssim L^2(\sqrt{s}\lambda + n_{p,s}^{-\frac{1}{2}})^2 + \frac{1}{n^{2/3}}. \tag{3.5}$$

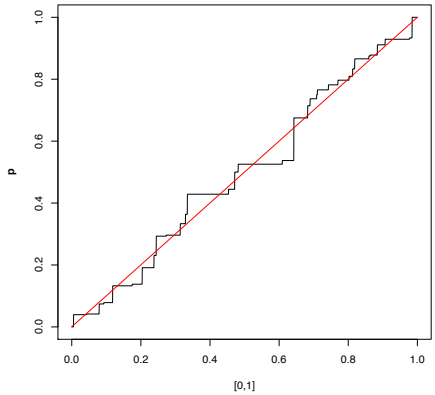
Note that (3.4) is satisfied when we set $\lambda = C\sqrt{\frac{\log p}{n}}$ for a sufficiently large constant C. Furthermore given that $\lambda = C\sqrt{\frac{\log p}{n}}$, (3.5) can be rewritten in the following way:

$$\frac{1}{n}\sum_{i=n+1}^{2n}(f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}))^2 \lesssim L^2 n_{p,s}^{-1} + \frac{1}{n^{2/3}}.$$

This shows the interplay between two rates: the nonparametric rate for estimating a monotone function – $O(\frac{1}{n^{2/3}})$, and the minimax rate for estimating the vector $\boldsymbol{\beta}^*$ – $O(n_{p,s}^{-1})$. Although this result holds only on half of the data, one can switch the two halves to obtain the same bound for the second half.



(a) Piecewise Constant $\mathbf{p}$ (4.1)



(b) General Monotone $\mathbf{p}$ (4.2)

Figure 1: We give two examples of models (1.2) with vectors $\mathbf{p}$ from (4.1) and (4.2). The true vector $\mathbf{p}$ is depicted with a red curve, while the estimate $\widehat{\mathbf{p}}$ is depicted with a black curve. We observe that the estimate in panel (a) traces more closely the red curve. This is to be expected in view of the results of Theorems 2.2 and 2.3

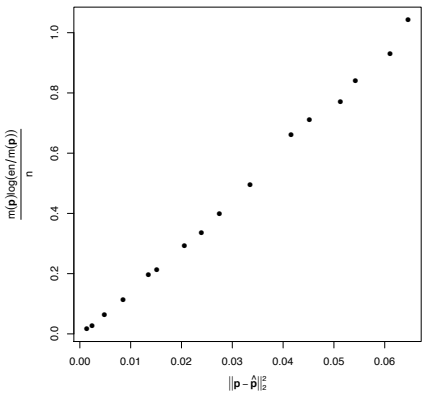
4 NUMERICAL EXPERIMENTS

In this section we briefly present results regarding the performance of the estimates developed in Sections 2 and 3. We first begin with results from Section 2. We consider the following two vectors

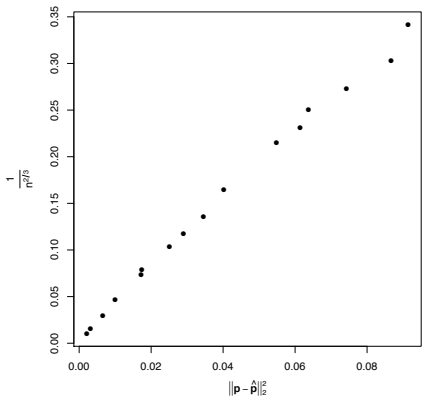
$$\mathbf{p} = \Big(\underbrace{\frac{1}{3}, \ldots, \frac{1}{3}}_{n/2}, \underbrace{\frac{2}{3}, \ldots, \frac{2}{3}}_{n/2}\Big), \tag{4.1}$$

and

$$\mathbf{p} : p_i = \frac{i}{n} \text{ for } i \in [n]. \tag{4.2}$$



(a) Piecewise Constant $\mathbf{p}$ (4.1)



(b) General Monotone $\mathbf{p}$ (4.2)

Figure 2: This figure shows the predicted estimation rates vs estimation rates averaged over 100 simulations for the two binary sequence models with probability vectors given in (4.1) and (4.2). Here, each point corresponds to sample size n which takes values in the set S_n. We observe near perfect linear alignment.

On Figure 1 we show two typical results of estimates using isotonic regression with $n = 2000$. We observe that on panel (a), the estimate (black curve) is closer to the truth (red curve) compared to the estimate in the panel (b). This is expected in view of the results of Theorems 2.2 and 2.3. In addition to these two examples we simulate models (1.2) for the two examples of $\mathbf{p}$ for a range of sample size values $n \in S_n := \{5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 45, 50, 100, 200, 500, 1000\}$ for a 100 simulations each. On Figure 2 we report the averaged value of $\|\mathbf{p} - \widehat{\mathbf{p}}\|_2^2$ over the 100 simulations, along with the theoretical value of the adaptive or non-adaptive rate (panel (a) and panel (b) respectively). We observe a near perfect linear alignment, confirming the findings of Theorems 2.2 and 2.3.

4.1 Application to Inverse Probability Weighting

In addition to the above examples in this subsection we will illustrate how one can apply isotonic regression to problems involving inverse probability weighting. Suppose we observe $(Y_i, X_i, A_i)_{i\in[n]}$ where $A_i \in \{0, 1\}$ indicates whether a patient was treated with a placebo or a new drug, $Y_i^{(a=0)} = f_0(X_i) + \varepsilon_i$ and $Y_i^{(a=1)} = f_1(X_i) + \varepsilon_i$ are the responses given that the patient was treated with placebo or the new drug resp., and $X_i \in \mathbb{R}$ is a covariate. For each patient exactly one of $Y^{(a=0)}$ or $Y^{(a=1)}$ is observed, and the goal is to estimate $\mathbb{E}Y^{(a=0)}$ and $\mathbb{E}Y^{(a=1)}$. According to our work, if the propensity score, $\mathbb{P}(A = 1|X = x)$, is an increasing function of x, one can use isotonic regression to find a nonparametric estimate $\widehat{\mathbb{P}}(A_i = 1|X_i)$ and then calculate

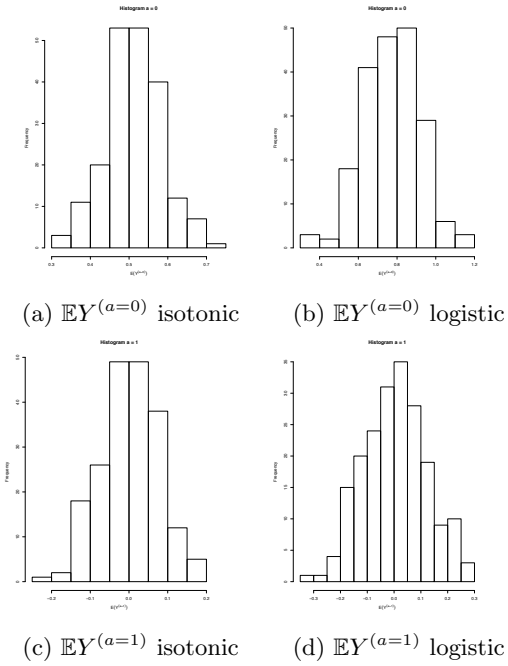
$$\sum_{i\in[n]} \frac{Y_i \mathbb{1}(A_i = 1)}{\widehat{\mathbb{P}}(A_i = 1|X_i)} \text{ and } \sum_{i\in[n]} \frac{Y_i \mathbb{1}(A_i = 0)}{\widehat{\mathbb{P}}(A_i = 0|X_i)},$$ [1]

to estimate $\mathbb{E}Y^{(a=1)}$ and $\mathbb{E}Y^{(a=0)}$. Below we show a numerical study using $n = 1000$, $X_i = \frac{i}{n}$, $\mathbb{P}(A_i = 1|X_i) = .1 + .8\mathbb{1}(i > 500)$, $f_0(x) = x$, $f_1(x) = 0$, $\varepsilon_i \sim N(0, 1)$. In this setting the true means are $\mathbb{E}Y^{(a=0)} = .5$ and $\mathbb{E}Y^{(a=1)} = 0$. In Figure 3 we plot 4 histograms of the treatment effect estimates over 200 repetitions – 2 of them are using isotonic regression and 2 of them are using logistic regression to model $\mathbb{P}(A = 1|X = x)$. We observe that for $\mathbb{E}Y^{(a=0)}$, the logistic regression approach gives heavily biased estimates centered at .8 while the isotonic regression correctly estimates the true mean of .5. In fact, the mean squared error of the logistic regression is 0.054 vs 0.005 of the isotonic regression. This is due to the failure of logistic regression to model the complicated distribution of $\mathbb{P}(A = 1|X = x)$. Note that in the above we can also use the SIM framework in cases when $X_i \in \mathbb{R}^p$ and $\mathbb{P}(A = 1|X_i) = g(X_i^\top \boldsymbol{\beta})$ for some increasing g.

4.2 Application to Binary Choice Models

We further provide a brief example of the performance of the two step procedure in binary SIM developed in Section 3. On Figure 4 we plot the results for model (1.3) using the logistic and probit links as defined in (3.2). We plot the curve $f(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})$ in red and the curve $\widehat{f}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})$ in black. We see that in both the logistic and in probit cases the estimates trace very well the corresponding target function. In this particular example the dimension $p = 2000$ the sparsity $s = 10$ and the sample size equals $n = 1000$ (so that the total sample size is $2n = 2000$). The unit vector $\boldsymbol{\beta}^*$ has equal

[1] Here we exclude observations for which $\widehat{\mathbb{P}}(A_i = 1|X_i)$ is very 0 or 1 for numerical stability.



(a) $\mathbb{E}Y^{(a=0)}$ isotonic (b) $\mathbb{E}Y^{(a=0)}$ logistic

(c) $\mathbb{E}Y^{(a=1)}$ isotonic (d) $\mathbb{E}Y^{(a=1)}$ logistic

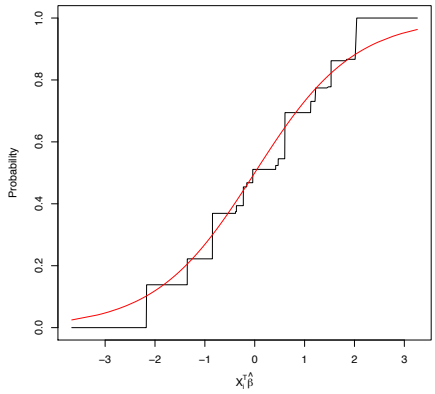
Figure 3: Four histograms of the estimates of $\mathbb{E}Y^{(a=0)}$ and $\mathbb{E}Y^{(a=1)}$ using logistic and isotonic regression respectively. The logistic regression approach gives heavily biased estimates for $\mathbb{E}Y^{(a=0)}$.

non-zero entries. The tuning parameter is selected via 10-fold cross validation.

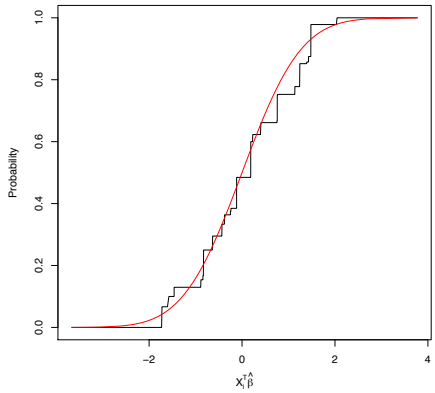
5 DISCUSSION

In this paper we presented results on the estimation of the monotone binary sequence model (1.2). We derived adaptive and non-adaptive rates of convergence, showing that akin to the Gaussian case, for probability vectors which are piecewise constant the isotonic regression has near parametric rates of convergence, whereas the rates of convergence for general probability vectors are of the order $\frac{1}{n^{2/3}}$. We used our results to develop a procedure for estimating f and $\boldsymbol{\beta}$ in a binary SIM with Gaussian design. We furthermore illustrated the success of our theory with simulated examples.

In addition to monotonicity one may be interested in other shape restrictions on the probability vector $\mathbf{p}$, such as convexity or unimodality. We believe that similar techniques to those we used in the present paper can be used to show that convexity and unimodality in the binary sequence model behave similarly to the Gaussian sequence model. We leave the details of this analysis to future work.



(a) Logit Link



(b) Probit Link

Figure 4: We give two examples of models (1.3) with logit $f(x) = f_{\text{logistic}}(x)$ (panel (a)) and probit $f(x) = f_{\text{probit}}(x)$ (panel (b)). In red is the true curve evaluated over the set $\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}$, i.e., $f(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})$ while the black curve is $\widehat{f}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}})$. In these examples we have set $p = 2000$, $s = 10$, $n = 1000$. The tuning parameter λ is selected via 10-fold cross validation. We see that the black curve traces well the red curve in both cases.

ACKNOWLEDGEMENTS

The author is grateful to the anonymous reviewers and area chair for multiple suggestions which greatly improved the presentation of this manuscript.

References

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, pages 641–647, 1955.

M. Banerjee and J. A. Wellner. Likelihood ratio tests for monotone functions. *Annals of Statistics*, pages 1699–1731, 2001.

P. C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.

P. C. Bellec and A. B. Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *Journal of Machine Learning Research*, 16: 1879–1892, 2015.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 16–21. IEEE, 2008.

S. Chatterjee, A. Guntuboyina, B. Sen, et al. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4): 1774–1800, 2015.

S. Chatterjee et al. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42 (6):2340–2381, 2014.

C. Gao, F. Han, and C.-H. Zhang. Minimax risk bounds for piecewise constant models. *arXiv preprint arXiv:1705.06386*, 2017.

P. Groeneboom and J. A. Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Birkhäuser, 2012.

A. Guntuboyina and B. Sen. Global risk bounds and adaptation in univariate convex regression. *Probability Theory and Related Fields*, 163(1-2):379–411, 2015.

A. Guntuboyina and B. Sen. Nonparametric shape-restricted regression. *arXiv preprint arXiv:1709.05707*, 2017.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Stat.*, 28(5):1302–1338, 2000. ISSN 0090-5364. doi: 10.1214/aos/1015957395. URL http://dx.doi.org/10.1214/aos/1015957395.

P. Mair, K. Hornik, and J. de Leeuw. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32(5):1–24, 2009.

P. Mörters and Y. Peres. *Brownian motion*, volume 30. Cambridge University Press, 2010.

M. Neykov, J. S. Liu, and T. Cai. L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *Journal of Machine Learning Research*, 17(87):1–37, 2016.

Y. Plan and R. Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1): 1–40, 2017.

T. Robertson, F. Wright, and R. Dykstra. Order restricted statistical inference. 1988.

C. Thrampoulidis, E. Abbasi, and B. Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems*, pages 3420–3428, 2015.

K. Tian, W. Kong, and G. Valiant. Learning populations of parameters. In *Advances in Neural Information Processing Systems*, pages 5780–5789, 2017.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

C.-H. Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002.