Estimation of Monotone Multi-Index Models

David Gamarnik∗
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139
gamarnik@mit.edu

Julia Gaudio†
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139
jgaudio@mit.edu

Abstract

In a multi-index model with \( k \) index vectors, the input variables are transformed by taking inner products with the index vectors. A transfer function \( f : \mathbb{R}^k \to \mathbb{R} \) is applied to these inner products to generate the output. Thus, multi-index models are a generalization of linear models. In this paper, we consider monotone multi-index models. Namely, the transfer function is assumed to be coordinate-wise monotone. The monotone multi-index model therefore generalizes both linear regression and isotonic regression, which is the estimation of a coordinate-wise monotone function. We consider the case of nonnegative index vectors. We provide an algorithm based on integer programming for the estimation of monotone multi-index models, and provide guarantees on the \( L_2 \) loss of the estimated function relative to the ground truth.

1 Introduction

Let \( \beta \) be a \( d \times k \) matrix, and let \( f \) be a function from \( \mathbb{R}^k \) to \( \mathbb{R} \). The model \( \mathbb{E}[Y|X] = f(\beta^T X) \) is known as a multi-index model. The columns of \( \beta \) are referred to as the index vectors and \( f \) is called a transfer function. Therefore, multi-index models generalize linear models. Typically, \( f \) is assumed to lie in a particular function class. In this paper, we assume that \( f \) is coordinate-wise monotone and satisfies a mild Lipschitz condition. We treat the case where the components of \( X \) are i.i.d., and the entries of \( \beta \) are nonnegative.

Supposing that the index vectors were known, the estimation of the function \( f \) would reduce to isotonic regression, which is the problem of estimating an unknown coordinate-wise monotone function. Monotone multi-index models (MMI) thereby additionally generalize isotonic regression. The setting where the transfer function is known is called the Generalized Index Model, a widely applicable statistical model [2]. We are therefore considering a much more challenging model.

We consider a high-dimensional setting, namely the dimension \( d \) is possibly much larger than the sample size, \( n \). We solve a sparse high-dimensional model; we assume that the number of index vectors (columns of \( \beta \)) is constant, and that \( \beta \) has a constant number \( s \) of nonzero rows. Finally, we assume that \( \beta \) is a nonnegative matrix, which is natural in many applications. For example, consider the following finance application. Suppose there are \( k \) future time periods, and \( d \) products. Let \( \beta(i, j) \) be the predicted monetary value of owning one unit of product \( i \) at a time \( j \). Given a vector \( x \) of product quantities, the value \( \beta^T x \) is a \( k \)-dimensional vector indicating the value of the products over the \( k \) time periods. Let \( f \) be a time-discounted measure of the overall value of the goods. Taking the example further, row sparsity would model an inventory restriction where one can store only \( s \) distinct types of goods.

∗http://web.mit.edu/gamarnik/www/home.html
†http://web.mit.edu/jgaudio/www/index.html

Preprint. Under review.
Work on index models has largely focused on the single-index model \((k = 1)\) (e.g. \([7, 9, 10, 13, 11]\)). In particular, \([13]\) provides the first provably efficient estimation algorithm for estimation of single-index models under monotonicity and Lipschitz assumptions. This work is further improved by \([11]\). To our knowledge, our paper is the first work done on estimation of multi-index models under the monotone Lipschitz model.

1.1 Notation

Let \(x\) be a vector in \(\mathbb{R}^d\). The vector \(p\)-norm \(\|x\|_p\) is defined as \(\|x\|_p^p \triangleq \sum_{i=1}^d x_i^p\). The \(\infty\)-norm is defined as \(\|x\|_{\infty} \triangleq \max_{i \in [d]} |x_i|\). Let \(M_{d,k}(r)\) be the set of \(d \times k\) matrices with each column having 2-norm at most \(r\). Similarly, let \(\overline{M}_{d,k}(r)\) be the set of \(d \times k\) matrices with each column having 2-norm equal to \(r\). Let \(O_{d,k}\) be the set of \(d \times k\) orthonormal matrices. Let \(P_k\) denote the set of \(k \times k\) rotation matrices, i.e. \(P_k = \{P \in \mathbb{R}^{k \times k} : P^T = P^{-1}, \text{det}(P) = 1\}\).

For a \(d \times k\) matrix \(M\), let \(M_{ij}^+ \triangleq \max\{M_{ij}, 0\}\), for \(i \in [d]\) and \(j \in [k]\), i.e. \(M^+\) is the matrix formed from \(M\) by replacing each negative entry by 0. Similarly, for a vector \(x \in \mathbb{R}^k\), let \(x^+\) denote the positive part of \(x\), i.e. \(x_i^+ = \max\{x_i, 0\}\). For a matrix \(M \in \mathbb{R}^{d \times k}\) and \(I \subseteq [d]\), let

\[
M(I)_{ij} = \begin{cases} 
M_{ij} & i \in I \\
0 & i \notin I.
\end{cases}
\]

In other words, the matrix \(M(I)\) is formed from \(M\) by zeroing all rows with index not belonging to \(I\). Similarly, for a vector \(x \in \mathbb{R}^d\), let \(x(I)_i = 1_{i \in I} x_i\). Note that \((M(I))^T x = M^T(x(I))\).

Let \(\| \cdot \|_p\) also denote the component-wise \(p\)-norm of a matrix, i.e. for a \(d \times k\) matrix \(M\), we have \(\|M\|_p^p = \sum_{i=1}^d \sum_{j=1}^k M_{ij}^p\). The Frobenius norm \(\|M\|_F\) is equal to \(\|M\|_2\) under this notation.

We say a function \(f : \mathbb{R}^k \rightarrow \mathbb{R}\) is \(l\)-Lipschitz if for every \(x, y \in \mathbb{R}^k\) it holds that

\[
|f(x) - f(y)| \leq l\|x - y\|_2.
\]

We say that \(f : \mathbb{R}^k \rightarrow \mathbb{R}\) is coordinate-wise monotone if for all \(x, y \in \mathbb{R}^k\) with \(x_i \leq y_i\) for each coordinate \(i\), it holds that \(f(x) \leq f(y)\). In other words, \(f\) is coordinate-wise monotone if it is monotone with respect to the Euclidean partial order. Fix \(b > 0\). Let \(C(b)\) be the set of coordinate-wise monotone functions \(f : \mathbb{R}^k \rightarrow [0, b]\), and let \(L_1(b)\) be the set of \(1\)-Lipschitz coordinate-wise monotone functions \(f : \mathbb{R}^k \rightarrow [0, b]\). Note that \(L_1(b) \subseteq C(b)\).

For a matrix \(\beta\) and function \(f\), write \((f \circ \beta)(x) \triangleq f(\beta^T x)\). Finally, let \(L(x, y, f) \triangleq (f(x) - y)^2\) be the loss function we consider.

1.2 The Model

We now describe the model. Some of the assumptions are carried from \([14]\). All parameters except the dimension \(d\) are considered constant.

Let \(\beta \in \overline{M}_{d,k}(r)\) be a \(d \times k\) matrix of rank \(k\), where each column has 2-norm equal to \(r\). Assume that \(\beta^*\) is \(s^*\)-row sparse, meaning that \(\beta^*\) has at most \(s^*\) nonzero rows. Let \(I^* \subseteq [d]\) be the set of non-zero rows of \(\beta^*\), so that \(\beta^*(I^*) = \beta^*\). Since \(\beta^*\) has full column rank, we can write \(\beta^* = Q^* R^*\) as its QR decomposition, where \(Q^* \in O_{d,k}\) and \(R^* \in \overline{M}_{k,d}(r)\) is invertible. We further assume that \(\beta^* \geq 0\) entrywise.

Let \(p_0\) be a twice-differentiable density supported on \(\mathcal{X} \subseteq R\). Let \(p^* = \max_{x \in \mathcal{X}} p_0(x)\). Further assume that \(\mathcal{X} \subseteq [-C, C]\). Let \(X \in \mathbb{R}^d\) be a random variable with density \(f_X(x) = \prod_{i=1}^d p_0(x_i)\). We additionally assume that \(\mathbb{E}[X] = 0\). This is without loss of generality, as we could treat the random variable \(X - \mathbb{E}[X]\) with support contained in the set \([-2C, 2C]^d\).

Let \(s_0(x) = \frac{p_0(x)}{p_0(0)}\) for \(x \in \mathcal{X}\). Let \(f^* \in L_1(b)\) be a twice-differentiable function. We assume that \(\mathbb{E}\left[\nabla^2 f^*(\beta^* T X)\right] > 0\), a restriction that ensures that estimation of \(\beta^*\) is information-theoretically feasible \([14]\). Let \(\rho_0\) be the smallest eigenvalue of \(\mathbb{E}\left[\nabla^2 f^*(\beta^* T X)\right]\). Note that since \(\beta^*\) has a constant number of columns and a constant number of nonzero rows, the value \(\rho_0\) is itself a constant.
The model is
\[ Y = (f^* \circ \beta^*)(X) + Z. \]  
(1)

Here \( Z \) is independent from \( X \) and satisfies \( \mathbb{E}[Z] = 0 \). We assume that \( |Z| \leq \eta \) almost surely so that \( Y \in \mathcal{Y} = [-\eta, b + \eta] \) almost surely. Let \( F(x, y) \) denote the joint density of \( X \) and \( Y \). We make a mild distribution assumption, which is that there exists \( \theta \) such that \( \mathbb{E}[80(X)^6] \leq \theta \) and \( \mathbb{E}[Y^n] \leq \theta \). Note that since \( Y \in [-\eta, b + \eta] \), then \( Y^6 \leq (b + \eta)^6 \) almost surely.

Given i.i.d. samples \((X_1, Y_1), \ldots, (X_n, Y_n)\) drawn from the model (1), our goal is to estimate the function \( f^* \circ \beta^* \), which is an element of the function class
\[
\mathcal{F}_{d,k} \triangleq \{ f \circ (I) : f \in \mathcal{L}_1(b), I \subset [d], |I| = s^*, \beta(I) \in \mathcal{M}_{d,k}(r) \}.
\]

**Proposition 1.** Let \( \mathcal{F}_{d,k} \triangleq \{ f \circ (I) : f \in \mathcal{L}_1(b), I \subset [d], |I| = s^*, \beta(I) \in \mathcal{M}_{d,k}(r) \} \). It holds that \( \mathcal{F}_{d,k} = \mathcal{F}_{d,k} \).

By Proposition 1 the model captures \( \beta^* \in \mathcal{M}_{k,k}(r) \subset \mathcal{M}_{k,k}(r) \) as well. Observe that for \( l > 0 \),
\[
f \circ \beta = f(lx) \circ \beta \frac{l}{I}.
\]

By this identity, the assumption that \( f^* \) is 1-Lipschitz and \( \beta^* \) has columns of norm \( r \) is without loss of generality; the assumption is equivalent to the assumption that \( f^* \) is \( 1 \)-Lipschitz and \( \beta^* \) has columns of norm \( r/l \).

### 1.3 Outline

We combine the results of two recent papers in order to design an algorithm for estimation in MMI models. [14] provide an algorithm for estimation of \( Q^* \) up to rotation given samples from the model (1). In other words, they find \( Q \) such that \( QP \approx Q^* \) for some rotation matrix \( P \). In Section 2 we summarize the approach of [14] to estimate the matrix \( Q^* \), up to rotation.

Informally, observe that if \( QP \approx Q^* \) and \( R \approx PR^* \), then \( QR \approx Q^* R^* \). Given a \( Q \) that approximates \( Q^* \) up to rotation, it remains to find \( R \in \mathcal{M}_{k,k}(r) \), an index set \( I \) of cardinality \( s^* \), as well as a function \( f \). Thus, the estimation of \( Q^* \) up to rotation reduces the high-dimensional estimation problem to a lower-dimensional problem.

Our approach is to form a collection of candidate \( k \times k \) matrices (Section 3). For each candidate matrix, we find the optimal index set and accompanying coordinate-wise monotone function (Section 4). We call the problem of finding the optimal index set \( I \) and coordinate-wise monotone function \( f \) the Sparse Matrix Isotonic Regression Problem. We extend the recent work of [5], who consider a related isotonic regression problem.

In Section 3 we tie together the results of the previous three sections in order to provide an algorithm for estimation in the high-dimensional monotone multi-index model. The algorithm finds a function of the form \( f \circ (QR)^+(I) \) minimizing the sample loss over the candidate matrices \( R \). Here \( I \) is an index set and \( f \) is a coordinate-wise monotone function obtained by solving the Sparse Matrix Isotonic Regression Problem. We give estimation guarantees for our algorithm in terms of \( L_2 \) loss. Finally, Section 5 outlines some future directions.

### 1.4 Contributions

Let \( \|f^* \circ Q^* R^* - g\|_2 \) denote the expected \( L_2 \) loss of a function \( g \) with respect to the ground truth:
\[
\|f^* \circ Q^* R^* - g\|_2^2 \triangleq \int_{x \in \mathbb{R}^d} [(f^* \circ Q^* R^*)(x) - g(x)]^2 f_X(x) dx.
\]

Let \( z(\epsilon_1, \epsilon_2, C) \triangleq 2\eta C \sqrt{k} (\epsilon_1 + \epsilon_2 r) + C^2 k (\epsilon_1 + \epsilon_2 r)^2 \). The main result of our paper is the following.

**Theorem 1.** Fix \( \epsilon > 0 \). Let \( \delta = \delta(\epsilon) \) be the solution to \( z(\delta, \delta, C) = \epsilon/2 \). Suppose \( d \geq \sqrt{2} \). Given \( n \) independent samples \((X_i, Y_i)_{i=1}^n\) from the model (1), there exists an algorithm that produces an estimator \( f_n \circ M_n^+(I_n) \) such that
\[
P \left( \|f_n \circ M_n^+(I_n) - f^* \circ Q^* R^*\|_2^2 \geq \epsilon \right) \leq \epsilon.
\]
whenever \( n \geq C_1 \log(d) + C_2 \), for constants \( C_1 \) and \( C_2 \) depending on \( C, b, s^*, p^*, k, \rho_0, \theta, \) and \( \eta \).
We summarize the work of [14], who estimate $Q^*$ up to rotation. The approach of [14] uses the second-order Stein condition. For $x \in \mathbb{R}^d$, let $T(x)$ be the $d \times d$ matrix defined as follows.

$$ T(x)_{ij} = \begin{cases} s_0(x_i)s_0(x_j) & i \neq j \\ s_0(x_i)^2 - s_0(i) & i = j. \end{cases} $$

[14] show the identity $\mathbb{E}[Y^\top T(X)] = Q^* D_0 Q^*$, where $D_0 = \mathbb{E} \left[ \nabla^2 f(\beta^* X) \right]$. Therefore, one can estimate $Q^*$ from the leading eigenvectors of the sample average of the quantity $Y^\top T(X)$. [14] use a robust estimator for $Y^\top T(X)$. For $\tau > 0$, define the truncated random variables

$$ \tilde{Y}_i \triangleq \text{sign}(Y_i) \cdot \min\{|Y_i|, \tau\} \quad \text{and} \quad \tilde{T}_{jk}(X_i) \triangleq \text{sign}(T_{jk}(X_i)) \cdot \min\{|T_{jk}(X_i)|, \tau^2\}. $$

The robust estimator is given by

$$ \tilde{\Sigma} = \tilde{\Sigma}(\tau) \triangleq \frac{1}{n} \sum_{i=1}^n \tilde{Y}_i \cdot \tilde{T}(X_i). $$

[14] propose the following approach to estimate $Q^*$ up to rotation.

**Algorithm 1** Estimation of $Q^*$ [14]

**Input:** Values $(X_1, Y_1), \ldots, (X_n, Y_n)$, $\tau > 0$, and $\lambda > 0$

**Output:** $\hat{Q} \in O_{d,k}$

1. Compute the estimator $\tilde{\Sigma}(\tau)$ using the samples $(X_i, Y_i)_{i=1}^n$.
2. Solve the following optimization problem.

$$ \begin{align*}
\max & \quad Tr(W^T \tilde{\Sigma}(\tau)) + \lambda||W||_1 \\
\text{s.t.} & \quad 0 \preceq W \preceq I_d \tag{2} \\
& \quad Tr(W) = k. \tag{3}
\end{align*} $$

3. Let $\hat{Q}$ be the matrix whose columns are the $k$ leading eigenvectors of $\hat{W}$.

**Theorem 2** (Adapted from Theorem 3.3 from [14]). Let $\tau = \left( \frac{3\rho \ln n}{\tau \log d} \right)^\frac{1}{2}$ and $\lambda = 10 \sqrt{\theta \log d / n}$. Under the assumptions of Section 1.2, with probability at least $1 - d^{-2}$, Algorithm 1 applied to samples $(X_i, Y_i)_{i=1}^n$, $\tau$, and $\lambda$ produces an estimator $\hat{Q}$ satisfying

$$ \inf_{P \in \mathcal{P}_k} ||\hat{Q} P - Q^*||_F \leq \frac{1}{\rho_0} 4 \sqrt{2} s^* \lambda. $$

Assuming that $d$ grows with $n$, Theorem 2 shows that with high probability as $n \to \infty$, the estimate of $Q^*$ is correct up to rotation, with error on the order of $\sqrt{\log(d)/n}$.

### 3 Construction of a Near-Net

Fix $\delta > 0$. We construct a random set of matrices $\mathcal{R}$ that will serve to approximate the set of $k \times k$ matrices with column norm $r$, with respect to the Frobenius norm. Given $\epsilon, \delta > 0$, we choose the cardinality of the set of approximating matrices so that a fixed matrix from the set $\bigcup_{k,k} \{r\}$ is $\epsilon$-close to some element of $\mathcal{R}$ with probability $1 - \delta$. For this reason, we call $\mathcal{R}$ a near-net. To construct $\mathcal{R}$, we first construct a random set of vectors $\mathcal{R}_0$ by choosing $N_0$ vectors from the uniform measure of all vectors of 2-norm $r$. In other words, each element from $\mathcal{R}_0$ is chosen from the uniform
measure on the $k$-dimensional sphere of radius $r$, denoted by $S^k_r$. We may sample uniformly using $k$ independent random variables $Z_1, \ldots, Z_k \sim \mathcal{N}(0, 1)$. The random vector
\[
\frac{1}{\sum_{i=1}^{k} Z^2_i} (Z_1, \ldots, Z_k)
\]
is uniformly distributed on the surface of $S^k_r$. Finally, we then construct the set $\mathcal{R}$ as the set of all matrices with columns belonging to $\mathcal{R}_0$. Then $|\mathcal{R}| = N_0^k$.

For a vector $x \in \mathbb{R}^k$ and $\epsilon > 0$, let $B(x, \epsilon) = \{ y : \|x - y\|_2 \leq \epsilon \}$ denote the ball of radius $\epsilon$ around $x$ with respect to the 2-norm.

**Lemma 1.** Let $\epsilon > 0$. Consider the near-net $\mathcal{R}(N_0)$ described above, and let $M \in \mathcal{M}_{k,k}(r)$ be a fixed matrix. With probability at least
\[
1 - k \left( 1 - \left| S^k_r \cap B \left( e_1, \frac{\epsilon}{\sqrt{k}} \right) \right| |S^k_r|^{-1} \right)^N_0 ,
\]there exists $\mathcal{R} \in \mathcal{R}(N_0)$ such that $\|M - \mathcal{R}\|_F \leq \epsilon$, where $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^k$ and $|A|$ denotes the measure of a set $A$.

**Remark 1.** As $N_0 \to \infty$, the probability (5) goes to 1.

Our random construction is simple to implement. While deterministic constructions are possible, they are much more complex (see [3]).

## 4 Sparse Matrix Isotonic Regression

Recently, Gamarnik and Gaudio introduced the Sparse Isotonic Regression model [5]. We now introduce a new related model, **Sparse Matrix Isotonic Regression**. We are given a $d \times k$ matrix $M$ with nonnegative entries as well as samples $(X_i, Y_i)_{i=1}^n$. For a given sparsity level $s \in \mathbb{N}$ and bound $b > 0$, our goal is to find a set $I \subset [d]$ with cardinality $s$, and a coordinate-wise monotone function $f : \mathbb{R}^k \to [0, b]$ minimizing $\sum_{i=1}^n (Y_i - (f \circ M(I))(X_i))^2$. Our approach is to estimate the function values at the points $X_1, \ldots, X_n$ and interpolate. We emphasize that we do not require the function $f$ to be 1-Lipschitz.

The Integer Programming Sparse Matrix Isotonic Regression algorithm finds the optimal index set and function values on a given set of points, given a matrix $M$ with nonnegative entries. Binary variables $v_i$ determine the index set $I$. The variables $F_i$ represent the estimated function values at data points $X_i$. Auxiliary variables $z_{ij}$ and $q_{ij,p}$ are used to model the monotonicity constraints. The function that is returned is an interpolation of the points $(M(I)^T X_i, F_i)_{i=1}^n$.

### Algorithm 2 Integer Programming Matrix Isotonic Regression

**Input:** Values $(X_1, Y_1), \ldots, (X_n, Y_n)$, sparsity level $s$, $M \geq 0 \in \mathbb{R}^{d \times k}$, $C > 0$, $b > 0$

**Output:** An index set $I \subset [d]$ satisfying $|I| = s$; a coordinate-wise monotone function $f : \mathbb{R}^k \to [0, b]$

1. Let $B = 2C \sum_{l=1}^d \sum_{p=1}^k M_{lp}$. Let
\[
\mu = \min \{ M_{lp} > 0 : l \in [d], p \in [k] \} \cdot \min_{i,j \in [n], i \neq j} |X_{di} - X_{dj}|,
\]

2. Solve the following optimization problem.
\[
\begin{align*}
\min_{v_i,F_i} & \sum_{i=1}^n (Y_i - F_i)^2 \\
\text{s.t.} & \sum_{l=1}^d v_l = s \\
& F_i - F_j \leq bz_{ij} \\
& \sum_{p=1}^k q_{ij,p} \geq z_{ij} \quad \forall i, j \in [n]
\end{align*}
\]
We note that it is possible to ensure that the function \( \hat{W} \) is \( \text{wise monotone} \), by modifying the optimization problem in Algorithm 2. However, the resulting integer program in Algorithm 2 has a convex objective and linear constraints. While integer programming is NP-hard in general, modern solvers achieve excellent performance on such problems.

We note that it is possible to ensure that the function \( \hat{f}_n \) be 1-Lipschitz in addition to coordinate-wise monotone, by modifying the optimization problem in Algorithm 2. However, the resulting optimization problem is an integer program with nonlinear constraints, a less tractable formulation. For further details, please see Section 8 in the supplementary material.

5 Estimation guarantees for the MMI Model

In this section, we provide estimation guarantees for the model \( Y = (f * \circ Q^* R^*)(X) + Z \). Let \( N \equiv 2n \) be the sample size. We use \( n \) samples for estimation of \( Q^* \) (up to rotation), obtaining a matrix \( Q_n \), and another \( n \) samples to obtain a matrix \( R_n \) a function \( f_n \), and index set \( I_n \). The final result is an estimated function \( \hat{f}_n \circ (Q_n R_n)^+(I_n) \).

We now outline the approach. First, by Theorem 2 the matrix \( Q_n \) obtained from the semidefinite programming approach satisfies \( \|Q_n P_n - Q^*\|_F \leq \frac{1}{\sqrt{n}} 4 \sqrt{2} s \lambda \) with probability at least \( 1 - d^{-2} \). We use this matrix \( Q_n \) to estimate \( R_n, f_n \) and \( I_n \), assuming that \( \|Q_n P_n - Q^*\|_F \leq \frac{1}{\sqrt{n}} 4 \sqrt{2} s \lambda \) for some unknown rotation matrix \( P_n \). The joint estimation of \( (R_n, f_n, I_n) \) is intractable; instead, we create a net of candidate matrices from the set \( \mathcal{M}_{k,k}(r) \). For each net element \( R \), we apply Algorithm 2 to find the optimal pair \( (f_R, I_R) \) minimizing the empirical loss \( \sum_{i=n+1}^{2n} L(X_i, Y_i, f \circ (Q_n R)^+(I)) \).

Finally, we output the best combination over the net elements.

Recall that \( f^* \circ \beta^* \in \mathcal{F}_{d,k}(r) \). While \( Q_n \in \mathcal{O}_{d,k} \) and \( R_n \in \mathcal{M}_{k,k}(r) \), the matrix \((Q_n R_n)^+(I_n)\) may not be an element of \( \mathcal{M}_{d,k}(r) \). Further, the estimated function \( f_n \) may not be 1-Lipschitz. Nevertheless, we are able to give an \( L_2 \) loss guarantee, as we will see in the proof of Theorem 3.

Algorithm 3 MMI Regression

Input: \( N_0 \in \mathbb{N}, \) values \( (X_1, Y_1), \ldots, (X_N, Y_N), C > 0, b > 0, \tau > 0, \lambda > 0 \)
Output: \( f_n \in \mathcal{L}(b), Q_n \in \mathcal{O}_{d,k}, R_n \in \mathcal{M}_{k,k}(r), \) and \( I_n \in [d]: |I_n| = s^* \)
1: Construct a random near-net \( \mathcal{R}(N_0) \).
2: Produce an estimate \( Q_n \) using Algorithm 2 applied to \( (X_i, Y_i)_{i=1}^n, \tau, \) and \( \lambda \).
3: for each \( R \in \mathcal{R} \) do
4: Apply Algorithm 2 to input \( (X_i, Y_i)_{i=n+1}^{2n}, s^*, (Q_n R)^+, C, \) and \( b \), obtaining the function \( f_R \) and index set \( I_R \).
5: end for
6: Return the tuple \((f_R, Q_n, R, I_R)\) with the smallest empirical loss.

The following result provides an upper bound on the error associated with the estimator from Algorithm 3. Our main result, Theorem 1 easily follows from Theorem 3.

Theorem 3. Let \( X \in \mathbb{R}^d \) be a random variable with independent entries of density \( p_0 \leq p^* \) and support contained within the set \([-C, C]^d \). Assume that \( f^*: \mathbb{R}^k \rightarrow [0, b] \) for \( b > 0 \). Fix \( n \). Let
\( \tau = \left( \frac{3\gamma n}{2\log d} \right)^{\frac{1}{6}} \) and \( \lambda = 10 \sqrt{\frac{\log d}{n}}. \) Let \( \epsilon > 0 \) and let \( \delta > 0 \) be such that \( \epsilon > z \left( \delta, \frac{1}{p_0} 4\sqrt{2}s^* \lambda, C \right). \)

Let \( (f_n, Q_n, R_n, I_n) \) be the result of applying Algorithm \( 2 \) on inputs \( N_0, (X_1, Y_1), \ldots, (X_{2n}, Y_{2n}), C, b, \tau, \) and \( \lambda. \) Let \( M_n = Q_n R_n. \) Then

\[
P \left( \| f_n \circ M_n^+ (I_n) - f^* \circ Q^* R^* \|_2^2 \geq \epsilon \right) \leq k \left( 1 - \left| S_{r_1} \right| \cap B \left( \epsilon_1, \frac{\delta}{\sqrt{r}} \right) \right) N_0 + \frac{1}{d^2} + 4 \left( \frac{d \epsilon}{s^*} \right)^k N_0 \exp \left( \frac{2 \log(2)b}{\alpha} + \frac{2 \log(4)p^* C^*}{\alpha} n \right),
\]

where \( \epsilon_0 = \epsilon - z \left( \delta, \frac{1}{p_0} 4\sqrt{2}s^* \lambda, C \right) \) and \( \alpha = \frac{1}{64} \epsilon_0 (b + \eta)^{-1}. \)

The following results are used in the proof of Theorem \( 3. \) Lemma \( 2 \) establishes a sensitivity result. The Lipschitz assumption on \( f^* \) is a key element in proving Lemma \( 2. \)

**Lemma 2.** Let \( X \in \mathbb{R}^d \) be a random variable with independent entries of density \( p_0 \leq p^* \) and support contained within the set \( [-C, C]^d. \) Suppose that \( R \in \mathcal{M}_{k,l}(r) \) satisfies \( \| PR - R \|_F \leq \epsilon_1. \) Suppose also that \( T \in O_{d,k} \) satisfies \( \| TP - Q \|_F \leq \epsilon_2 \) for some rotation matrix \( P \in \mathcal{P}_{k,l}. \) Then

\[
\int L(x, y, f^* \circ (TR^*) \cap (I^*)) dF(x, y) - \int L(x, y, f^* \circ Q^* R^*) dF(x, y) \leq z(\epsilon_1, \epsilon_2, C).
\]

**Lemma 3.** Let \( g \) be any function from \( \mathbb{R}^k \) to \( \mathbb{R}. \) Then

\[
\| g - f^* \circ Q^* R^* \|_2^2 = \int L(x, y, g) dF(x, y) - \int L(x, y, f^* \circ Q^* R^*) dF(x, y).
\]

Fix \( b > 0. \) For \( T \in O_{d,k}, R \in \mathcal{M}_{k,l}(r), \) and \( I \subset [d] \) with \( |I| = s^* \), let

\[
\mathcal{G}(T, R, I) = \{ f \circ (TR)^+ (I) : f \in C(b) \} \text{ and } \mathcal{G}(T, R) \triangleq \cup_{R \in \mathcal{R}} \cup_{I \subset [d], |I| = s^*} \mathcal{G}(T, R, I).
\]

We see that Algorithm \( 3 \) optimizes the empirical loss over functions in \( \mathcal{G}(Q_n, \mathcal{R}). \) We follow a VC entropy approach to give an \( L_2 \) loss bound for the function \( f_n \circ M_n^+ (I_n) \) estimated by Algorithm \( 3. \)

**Definition 1.** Let \( \mathcal{F} \) be a class of functions from \( \mathbb{R}^d \) to \( \mathbb{R}. \) Given

\[
(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R},
\]

let

\[
L_{\mathcal{F}} \triangleq \{ (L(x_1, y_1), \ldots, L(x_n, y_n), f) : f \in \mathcal{F} \}.
\]

In other words, \( L_{\mathcal{F}} \) is the set of loss vectors formed by ranging over functions in the class \( \mathcal{F}. \) Let \( N_{\mathcal{F}} ((x_1, y_1), \ldots, (x_n, y_n), \epsilon) \) denote the size of the smallest \( \epsilon \)-net for \( L_{\mathcal{F}} ((x_1, y_1), \ldots, (x_n, y_n)), \) with respect to the \( \infty \)-norm. In other words, for every \( u \in L_{\mathcal{F}} ((x_1, y_1), \ldots, (x_n, y_n)) \), there exists \( v \in N_{\mathcal{F}} ((x_1, y_1), \ldots, (x_n, y_n)) \) such that \( \| u - v \|_\infty \leq \epsilon. \) Finally, let

\[
N_{\mathcal{F}}(\epsilon, n) \triangleq \mathbb{E}_{X, Y} [N_{\mathcal{F}} ((X_1, Y_1), \ldots, (X_n, Y_n), \epsilon)]
\]

be the expected size of the net, where the expectation is over independent samples drawn from the distribution \( F(x, y) \) defined above.

Lemmas 4 and 5 together provide a probabilistic bound on the difference between expected loss and empirical loss for functions in the class \( \mathcal{G}(T, \mathcal{R}). \) The nonnegative matrix assumption is crucial for the proof of Lemma 5.

**Lemma 4.** Let \( T \in O_{d,k} \) and let \( \delta > 0. \) Let \( \mathcal{R} \subset \mathcal{M}_{k,l}(r). \) For \( \epsilon > 0, \)

\[
P \left( \sup_{h \in \mathcal{G}(T, \mathcal{R})} \left| \int L(x, y, h) dF(x, y) - \frac{1}{n} \sum_{i=1}^n L(X_i, Y_i, h) \right| \geq \epsilon \right) \leq 4 N_{\mathcal{G}(T, \mathcal{R})} \left( \frac{\epsilon}{16}, n \right) \exp \left( -\frac{\epsilon^2 n}{128b^2} \right).
\]

**Lemma 5.** Under the assumptions of Lemma 4, it holds that

\[
N_{\mathcal{G}(T, \mathcal{R})} (\epsilon, n) \leq \left( \frac{d}{s^*} \right)^k N_0 \exp \left( \frac{2 \log(2)b}{\alpha} + \frac{2 \log(4)p^* C^*}{\alpha} n \right),
\]

where \( \alpha = \frac{1}{2} \epsilon (b + \eta)^{-1}. \)
With these results stated, we are now ready to prove Theorem 3.

**Proof of Theorem 3.** With probability at least $1 - d^{-2}$, the matrix $Q_n$ satisfies

$$
\|Q_n P_n - Q^*\|_F \leq \frac{1}{\rho_0} 4\sqrt{2}s^*\lambda
$$

for some rotation matrix $P_n$ (Theorem 2). For the remainder, we condition on this property of $Q_n$, since we use this matrix on an independent batch of samples $(X_i, Y_i)_{i=n+1}^{2n}$. By Lemma 3,

$$
\|f_n \circ M_n^+(I_n) - f^* \circ Q^* R^*\|_2^2 = \int L(x, y, f_n \circ M_n^+(I_n))dF(x, y) - \int L(x, y, f^* \circ Q^* R^*)dF(x, y).
$$

Recall that $f^*$ is the set of non-zero rows of $\beta^*$. Let $E$ be the event that the near-net $R$ contains an element $\overline{R} \in \mathcal{R}$ such that $\|PR^* - \overline{R}\|_F \leq \delta$. Conditioned on $E$, let $\overline{R}$ be the (random) matrix that is $\delta$-close to $PR^*$. By Proposition 2, the function $f_n \circ M_n^+(I_n)$ is optimal over the samples. Therefore, $\sum_{i=n+1}^{2n} L(X_i, Y_i, f_n \circ M_n^+(I_n)) \leq \sum_{i=n+1}^{2n} L(X_i, Y_i, f^* \circ (Q_n \overline{R})^+(I^*))$. We have

$$
\|f_n \circ M_n^+(I_n) - f^* \circ Q^* R^*\|_2^2 \leq \int L(x, y, f_n \circ M_n^+(I_n))dF(x, y) - \frac{1}{n} \sum_{i=n+1}^{2n} L(X_i, Y_i, f_n \circ M_n^+(I_n)) + \int L(f^* \circ (Q_n \overline{R})^+(I^*))dF(x, y) - \int L(x, y, f^* \circ Q^* R^*)dF(x, y).
$$

By Lemma 2 applied to $T = Q_n$, $\epsilon_1 = \delta$, and $\epsilon_2 = \frac{1}{\rho_0} 4\sqrt{2}s^*\lambda,$

$$
\int L(f^* \circ (Q_n \overline{R})^+(I^*))dF(x, y) - \int L(x, y, f^* \circ Q^* R^*)dF(x, y) \leq z \left( \delta, \frac{1}{\rho_0} 4\sqrt{2}s^*\lambda, C \right).
$$

Therefore,

$$
\mathbb{P}\left( \|f_n \circ M_n^+(I_n) - f^* \circ Q^* R^*\|_2^2 \geq \epsilon \mid E \right) \leq \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} L(X_i, Y_i, f_n \circ M_n^+(I_n)) \right) + \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} L(X_i, Y_i, f^* \circ (Q_n \overline{R})^+(I^*)) - \frac{1}{n} \sum_{i=1}^{n} L(X_i, Y_i, h) \mid E \right).
$$

Observe that the functions $f_n \circ M_n^+(I_n)$ and $f^* \circ (Q_n \overline{R})^+(I^*)$ are elements of $G(Q_n, \mathcal{R})$. Since the event $E$ is independent from the samples $(X_i, Y_i)$, we apply Lemmas 3 and 5

$$
\mathbb{P}\left( \|f_n \circ M_n^+(I_n) - f^* \circ Q^* R^*\|_2^2 \geq \epsilon \mid E \right) \leq \mathbb{P}\left( \sup_{h \in G(Q_n, \mathcal{R})} \left| \frac{1}{n} \sum_{i=1}^{n} L(X_i, Y_i, h) \right| \geq \frac{\epsilon_0}{2} \mid E \right) \leq 4 \left( \frac{d}{s^*} \right)|\mathcal{R}| \exp \left[ \left( \frac{2\log(2)b}{\alpha} + 2\pi (4p^* C)^\star \right)n^{-\frac{\epsilon_0^2}{2\pi}} - \frac{\epsilon_0^2 n}{2^9 b^2} \right],
$$

where $\alpha = \frac{1}{64} \epsilon_0 (b + \eta)^{-1}$. The result follows by Lemma 1.

$$
\mathbb{P}(E^c) \leq k \left( 1 - \left| S_{r^k-1} \cap B \left( e_1 \frac{\epsilon}{\sqrt{k}} \right) \right| \left| S_{r^k-1} \right|^{-1} \right) N_0.
$$
6 Conclusion

In this paper, we have provided an estimation algorithm for multi-index models with a coordinate-wise monotone transfer function. Our algorithm enables future work on wide-ranging applications naturally modeled as a monotone multi-index model. Promising future directions include finding an efficient method for the sparse matrix isotonic regression problem, dropping the nonnegativity assumption of the index vectors, as well as studying multi-index models with other classes of transfer functions. Studying the non-sparse setting would be interesting as well, and would present several challenging technical hurdles.

References

[1] Gleb Beliakov. Monotonicity preserving approximation of multivariate scattered data. *BIT Numerical Mathematics*, 45(4):653–677, 2005.
[2] Annette J. Dobson and Adrian G. Barnett. *An introduction to generalized linear models*. CRC Press, 2018.
[3] Ilya Dumer. Covering spheres with spheres. *Discrete & Computational Geometry*, 38(4):665–679, 2007.
[4] David Gamarnik. Efficient learning of monotone concepts via quadratic optimization. In *COLT*, 1999.
[5] David Gamarnik and Julia Gaudio. Sparse high-dimensional isotonic regression. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.
[6] Dimitris Bertsimas David Gamarnik and John N. Tsitsiklis. Estimation of time-varying parameters in statistical models: an optimization approach. *Machine Learning*, 35(3):225–245, 1999.
[7] Wolfgang Härdle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The Annals of Statistics*, pages 157–178, 1993.
[8] David Haussler. Overview of the Probably Approximately Correct (PAC) learning framework. *https://hausslergenomics.ucsc.edu/wp-content/uploads/2017/08/smo_0.pdf*, 1995.
[9] Joel L. Horowitz and Wolfgang Härdle. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640, 1996.
[10] Hidehiko Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Technical report, Center for Economic Research, Department of Economics, University of Minnesota, 1991.
[11] Sham M. Kakade, Adam Tauman Kalai, Varun Kanade, and Ohad Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems 24 (NeurIPS 2011)*, 2011.
[12] Guy Moshkovitz and Asaf Shapira. Ramsey theory, integer partitions and a new proof of the Erdős–Szekeres theorem. *Advances in Mathematics*, 262:1107–1129, 2014.
[13] Adam Tauman Kalai and Ravi Sastry. The Isotron algorithm: High-dimensional isotonic regression. *Conference on Learning Theory*, 2009.
[14] Zhuoran Yang, Krishna Balasubramanian, Zhaoran Wang, and Han Liu. Learning non-Gaussian multi-index model via a second-order Stein’s method. *Advances in Neural Information Processing Systems*, 30:6097–6106, 2017.
7 Deferred proofs

Proof of Proposition[7] Clearly $\mathcal{F}_{d,k} \subseteq \mathcal{F}_{d,k}$. It remains to show that $\mathcal{F}_{d,k} \subseteq \mathcal{F}_{d,k}$. Let $g \in \mathcal{F}_{d,k}$, where $g = f \circ \beta(I)$. We will show that $g \in \mathcal{F}_{d,k}$.

Suppose the $i$th column of $\beta(I)$ has norm $t < r$. Let $\overline{\beta}$ be equal to $\beta(I)$, with the $i$th column scaled by a factor of $\frac{t}{r}$, so that the $i$th column of $\overline{\beta}$ has norm $r$. Note that $\overline{\beta} = \beta(I)$. Next, define the function $\overline{f}$ by $\overline{f}(x) = f(x_1, \ldots, \frac{t}{r}x_i, \ldots, x_k)$. Observe that $g = \overline{f} \circ \overline{\beta}$. We verify the monotonicity property for $\overline{f}$. Let $x \preceq y$. Then also $(x_1, \ldots, \frac{t}{r}x_i, \ldots, x_k) \preceq (y_1, \ldots, \frac{t}{r}y_i, \ldots, y_k)$, and we have

$$\overline{f}(x) = f\left(x_1, \ldots, \frac{t}{r}x_i, \ldots, x_k\right) \preceq f\left(y_1, \ldots, \frac{t}{r}y_i, \ldots, y_k\right) = \overline{f}(y).$$

It remains to show that $\overline{f}$ is 1-Lipschitz. Let $x, y \in \mathbb{R}^k$. By the Lipschitz condition applied to $f$,

$$-1 \leq \frac{\overline{f}(x) - \overline{f}(y)}{\|x - y\|_2} \leq 1 \implies -1 \leq \overline{f}(x) - \overline{f}(y) \leq 1.$$

This shows that $\overline{f}$ is 1-Lipschitz. Repeating this argument for each column of $\beta(I)$, we conclude that $g \in \mathcal{F}_{d,k}$. \hfill $\square$

To prove Lemma[6] we use the following helper lemma.

Lemma 6. Consider the near-net $R_0$ with $N_0$ described above, and let $v$ be a fixed vector of norm $r$. With probability

$$1 - \left(1 - \frac{1}{|S_k^d \cap B(e_1, \delta)|} \right)^{N_0},$$

there exists $\overline{v} \in R_0$ such that $\|v - \overline{v}\|_2 \leq \delta$.

Proof. Let $u$ be distributed uniformly at random on the surface of $S_k^{d-1}$. Then

$$\mathbb{P}\left(\|v - u\|_2 \leq \delta\right) = \mathbb{P}\left(\{u \in B(v, \delta)\}\right)$$

$$= \frac{1}{|S_k^{d-1} \cap B(v, \delta)|} \frac{1}{|S_k^d|}$$

$$= \frac{1}{|S_k^{d-1} \cap B(e_1, \delta)|} \frac{1}{|S_k^d|}.$$

Therefore,

$$\mathbb{P}\left(\|v - u\|_2 > \delta\right) = 1 - \frac{1}{|S_k^d \cap B(e_1, \delta)|} \frac{1}{|S_k^d|}.$$

We conclude that the probability that there exists an element of $R_0$ that is $\delta$-close to $v$ is equal to

$$1 - \left(1 - \frac{1}{|S_k^d \cap B(e_1, \delta)|} \frac{1}{|S_k^d|}\right)^{N_0}. \hfill \square$$

Proof of Lemma[7] Let $E$ be the event that there exists $\overline{R} \in \mathcal{R}$ such that $\|M - \overline{R}\|_F \leq \epsilon$. We need to lower bound the probability of the event $E$. Let $\{M_i\}_{i=1}^k$ denote the columns of $M$. For $i \in [k]$, let $E_i$ be the event that there exists $\overline{u}_i \in \mathcal{R}_0$ such that $\|M_i - \overline{u}_i\|_2 \leq \frac{\epsilon}{\sqrt{k}}$. We claim that $\mathbb{P}\left(\cap_{i=1}^k E_i\right) \leq \mathbb{P}(E)$. Indeed, suppose that the event $E_i$ occurs for each $i$. Let $\overline{R} \in \mathcal{R}$ be the matrix with columns $\{\overline{u}_i\}_{i=1}^k$. Then

$$\|M - \overline{R}\|_F^2 = \sum_{i=1}^k \|M_i - \overline{u}_i\|_2^2 \leq k \left(\frac{\epsilon}{\sqrt{k}}\right)^2 = \epsilon^2,$$

and so $\|M - \overline{R}\|_F \leq \epsilon$. This shows that $\mathbb{P}\left(\cap_{i=1}^k E_i\right) \leq \mathbb{P}(E)$. We also have

$$\mathbb{P}(E^c) \leq \mathbb{P}\left(\cup_{i=1}^k E_i^c\right) \leq \sum_{i=1}^k \mathbb{P}(E_i^c),$$

so that $\mathbb{P}(E) \geq 1 - \sum_{i=1}^k \mathbb{P}(E_i^c)$.
For each $i \in [k]$, Lemma 5 shows that there exists $\pi_i \in \mathcal{R}_0$ such that $\|M_i - \pi_i\|_2 \leq \epsilon/\sqrt{k}$, with probability
\[
1 - \left(1 - \left|S_i^b \cap B \left(e_1, \frac{\epsilon}{\sqrt{k}}\right)\right| \left|S_i^b\right|^{-1}\right)^{N_0}.
\]
Therefore,
\[
\mathbb{P}(E_i^1) \leq 1 - \left(1 - \left|S_i^b \cap B \left(e_1, \frac{\epsilon}{\sqrt{k}}\right)\right| \left|S_i^b\right|^{-1}\right)^{N_0}.
\]

**Proof of Proposition** 2. We first show that the constraints enforce the monotonicity requirement. Let $I = \{i : v_i = 1\}$. Consider two samples $(X_i, Y_i)$ and $(X_j, Y_j)$. The monotonicity requirement is
\[
(M(I))^T X_i \leq (M(I))^T X_j \quad \implies \quad F_i \leq F_j.
\]
The contrapositive of this statement is
\[
F_i > F_j \quad \implies \quad \exists p \in [k] : ((M(I))^T X_i)_p > ((M(I))^T X_j)_p.
\]
The optimization encodes the contrapositive statement, as follows. There are two cases: either $F_i > F_j$ or $F_i \leq F_j$. We must ensure that if $F_i > F_j$ holds, then the implication in (11) is satisfied. We must also verify that no additional constraints are introduced if $F_i \leq F_j$.

Suppose $F_i > F_j$. Then $z_{ij} = 1$ by Constraint (5). By Constraint (9), at least one of the $q_{ijp}$ variables must be equal to 1. Then by Constraints (10), we have
\[
\sum_{i=1}^d q_{ip}(X_{il} - X_{jl}) > 0 \iff ((M(I))^T X_i)_p > ((M(I))^T X_j)_p,
\]
for at least one $p \in [k]$, due to the choice of $\mu$. Next suppose $F_i \leq F_j$. Then $z_{ij}$ is free to equal zero, and all the $q_{ijp}$ values may be set to zero as well. By the choice of $B$, Constraint (10) is then non-binding.

The objective minimizes the loss on the samples. Finally, we claim that the choice of $\hat{f}_n$ is a monotone interpolation. First, $\hat{f}_n(M(I_n))^T X_i = F_i$, so that $\hat{f}_n$ interpolates. Next, observe that $x \leq y \implies \hat{f}_n(x) \leq \hat{f}_n(y)$. Also, $\hat{f}_n : \mathbb{R} \to [0, b]$, by construction.

**Proof of Lemma** 2. Fix $(x, y)$. We have
\[
L(x, y, f^* \circ (TR)^+(I^*))dF(x, y) - L(x, y, f^* \circ Q^* R^*)
\]
\[
= (y - (f^* \circ (TR)^+(I^*))(x))^2 - (y - (f^* \circ Q^* R^*))(x))^2
\]
\[
= ((f^* \circ (TR)^+(I^*))(x) - (f^* \circ Q^* R^*)(x)) \cdot (f^* \circ (TR)^+(I^*))(x) + (f^* \circ Q^* R^*)(x) - 2y
\]
\[
\leq |(f^* \circ Q^* R^*)(x) - (f^* \circ (TR)^+(I^*))(x)| \cdot |(f^* \circ Q^* R^*)(x) + (f^* \circ (TR)^+(I^*))(x) - 2y|
\]
(12)
Since $f^* \in L_1(b)$, it holds that
\[
|(f^* \circ Q^* R^*)(x) - (f^* \circ (TR)^+(I^*))(x)| \leq \|Q^* R^* - (TR)^+(I^*)\|_2
\]
\[
= \|Q^* R^* - (TR)^+(I^*)\|_2.
\]
For the second factor in the bound (12), we have
\[
|(f^* \circ Q^* R^*)(x) + (f^* \circ (TR)^+(I^*))(x) - 2y|
\]
\[
\leq 2|(f^* \circ Q^* R^*)(x) - y| + |(f^* \circ (TR)^+(I^*))(x) - (f^* \circ Q^* R^*)(x)|
\]
\[
\leq 2|f^* \circ Q^* R^*)(x) - y| + \|Q^* R^* - (TR)^+(I^*)\|_2.
\]
Let $A = Q^* R^* - (TR)^+(I^*)$. Substituting into (12), we have
\[
\int L(x, y, f^* \circ (TR)^+)dF(x, y) - \int L(x, y, f^* \circ Q^* R^*)dF(x, y)
\]
\[
\begin{align*}
\leq & \int \|A^T x\|_2^2 (2 \|f^* \circ Q^* R^*\|_F (x) - y) + \|A^T x\|_2^2 \, dF(x, y) \\
\leq & \int \|A^T x\|_2^2 (2\eta + \|A^T x\|_2^2) \, dF(x, y) \\
= & \int \|A^T x\|_2^2 (2\eta + \|A^T x\|_2^2) \, dF(x) \\
= & 2\eta \mathbb{E} \left[ \|A^T X\|_2^2 \right] + \mathbb{E} \left[ \|A^T X\|_2^2 \right] \\
\leq & 2\eta \sqrt{\mathbb{E} \left[ \|A^T x\|_2^2 \right] + \mathbb{E} \left[ \|A^T x\|_2^2 \right]} ,
\end{align*}
\]
where the last inequality follows from Jensen’s inequality. We now evaluate the expectation.

\[
\mathbb{E} \left[ \|A^T X\|_2^2 \right] = \mathbb{E} \left[ \sum_{j=1}^{k} \left( \sum_{i=1}^{d} A_{ij} X_i \right)^2 \right] = \sum_{j=1}^{k} \sum_{i=1}^{d} A_{ij}^2 \mathbb{E} \left[ X_i^2 \right] \leq C^2 \sum_{j=1}^{k} \sum_{i=1}^{d} A_{ij}^2 = C^2 \|A\|_F^2 .
\]

Using the fact that \(Q^* R^* \geq 0\) and \(Q^* R^* = (Q^* R^*)(I^*)\), we have

\[
\|A\|_F = \|Q^* R^* - (TR)^+ (I^*)\|_F \\
\leq \|Q^* R^* - TR\|_F \\
\leq \|Q^* R^* - TP R^*\|_F + \|TP R^* - TR\|_F \\
= \|Q^* - TP\|_F R^* \|_F + \|TP\|_F R^* - R\|_F \\
\leq \epsilon_2 R^* \|_F + \epsilon_1 \|T\|_F \\
= \epsilon_2 \sqrt{kr^2 + \epsilon_1 \sqrt{k}} \\
= \sqrt{k} (\epsilon_1 + \epsilon_2 r) .
\]

Therefore,

\[
\mathbb{E} \left[ \|A^T X\|_2^2 \right] \leq C^2 \|A\|_F^2 \leq C^2 k (\epsilon_1 + \epsilon_2 r)^2 .
\]

Substituting into (13), we conclude

\[
\int L(x, y, f^* \circ (TR)^+) \, dF(x, y) - \int L(x, y, f^* \circ Q^* R^*) \, dF(x, y) \\
\leq 2\eta \sqrt{k} (\epsilon_1 + \epsilon_2 r) + C^2 k (\epsilon_1 + \epsilon_2 r)^2 = z(\epsilon_1, \epsilon_2, C).
\]

\[\square\]

**Proof of Lemma** 3 A similar result appears in [6]. We include the proof for completeness.

\[
\int L(x, y, g) \, dF(x, y) \\
= \int (g(x) - y^2) \, dF(x, y) \\
= \int (g(x) - (f^* \circ Q^* R^*)(x) + (f^* \circ Q^* R^*)(x) - y^2) \, dF(x, y) \\
= \int (g(x) - (f^* \circ Q^* R^*)(x))^2 + ((f^* \circ Q^* R^*)(x) - y)^2 + 2(g(x) - (f^* \circ Q^* R^*)(x))((f^* \circ Q^* R^*)(x) - y) \, dF(x, y) \\
= \|g - f^* \circ Q^* R^*\|_2^2 + \int L(x, y, f^* \circ Q^* R^*) \, dF(x, y) + 2\mathbb{E} \left[ (g(X) - (f^* \circ Q^* R^*)(X))((f^* \circ Q^* R^*)(X) - Y) \right] .
\]

Since \(\mathbb{E}[Y \mid X] = (f^* \circ Q^* R^*)(x)\), we have

\[
\mathbb{E} \left[ (g(X) - (f^* \circ Q^* R^*)(X))((f^* \circ Q^* R^*)(X) - Y) \right] = 0 .
\]

Rearranging completes the proof. \[\square\]
We now work towards a proof of Lemma 5. Recall \( \alpha = \alpha(\varepsilon) = \frac{1}{\varepsilon}(b + \eta)^{-1} \) and let \( S = \{0, \alpha, 2\alpha, \ldots, \left\lceil \frac{b}{\alpha} \right\rceil \} \) be a discretization of the range \([0, b]\). For \( f \in C(b) \), let

\[
g_f(x) = \max\{q \in S : q \leq f(x)\}.
\]

In other words, the function \( g_f \) is formed by rounding each value down to the nearest increment in \( S \). Fix \( I \subset [d] \) with \( |I| = s^* \). Recall the definition \( G(T, R, I) = \{f \circ (TR)^+(I) : f \in C(b)\} \), and let \( H(T, R, I) = \{g_f \circ (TR)^+(I) : f \in C(b)\} \). Proposition 3 will show that the set \( L_{H(Q,R,I)}((x_1, y_1), \ldots, (x_n, y_n)) \) is an \( \varepsilon \)-net for the set \( L_{G(Q,R,I)}((x_1, y_1), \ldots, (x_n, y_n)) \). Next, we relate the cardinality of \( L_{H(Q,R,I)}((x_1, y_1), \ldots, (x_n, y_n)) \) to a labeling number.

**Definition 2** (Labeling Number [4]). For a sequence of points \( x_1, \ldots, x_n \in \mathbb{R}^k \) and a positive integer \( m \), the labeling number \( \Lambda(x_1, \ldots, x_n; m) \) is the number of functions \( \phi : \{x_1, \ldots, x_n\} \to \{1, 2, \ldots, m\} \) such that \( \phi(x_i) \leq \phi(x_j) \) whenever \( x_i \preceq x_j \) for \( i, j \in \{1, \ldots, n\} \).

Let \( M = (QR)^+ \). Observe that the cardinality of \( L_{H(Q,R,I)}((x_1, y_1), \ldots, (x_n, y_n)) \) is upper-bounded by the labeling number of the set \( \{M^Tx_1, \ldots, M^Tx_n\} \) with \( \left\lceil \frac{b}{\alpha} \right\rceil - 1 \) labels. Therefore, the value of \( N_{H(Q,R,I)}(\varepsilon, n) \) is upper-bounded by the expected labeling number of the set \( \{M^Tx_1, \ldots, M^Tx_n\} \) with \( \left\lceil \frac{b}{\alpha} \right\rceil - 1 \) labels.

We therefore need to determine an upper bound on the expected labeling number of the set \( \{M^Tx_1, \ldots, M^Tx_n\} \) with \( \left\lceil \frac{b}{\alpha} \right\rceil - 1 \) labels. Let \( \pi(I) \) be the vector formed from the entries of \( x \) that are indexed by the set \( I \). We will first show that the labeling number of the set \( \{M^Tx_1, \ldots, M^Tx_n\} \) is upper-bounded by the labeling number of the set \( \{\pi(I), \ldots, \pi(I)\} \) with the same number of labels. Observe that the points \( \{\pi(I)\}_{i=1}^n \) have dimension \( s^* \). We will then analyze the expected labeling number of a sequence of random variables \( \{W_1, \ldots, W_n\} \) that are of dimension \( s^* \).

The following proposition will be used to establish the net property.

**Proposition 3.** Let \( T \in O_{d,k}, R \in \mathbb{N}_{k,k}(r), \) and \( I \subset [d] \). Fix \( f \circ (TR)^+(I) \in G(T, R, I) \) and the accompanying \( g_f \circ (TR)^+ \in H(T, R, I) \). Let \( x \in \mathcal{X} \) and \( y \in \mathcal{Y} \). Then

\[
|L(x, y, f \circ (TR)^+(I)) - L(x, y, g_f \circ (TR)^+(I))| \leq \varepsilon.
\]

**Proof.** Let \( M = (TR)^+ \).

\[
|L(x, y, f \circ (TR)^+(I)) - L(x, y, g_f \circ (TR)^+(I))| = |L(x, y, f \circ M(I)) - L(x, y, g_f \circ M(I))| = |((f \circ M(I))(x) - y)^2 - ((g_f \circ M(I))(x) - y)^2| = |(f \circ M(I))^2(x) - (g_f \circ M(I))^2(x) - 2y((f \circ M(I))(x) - (g_f \circ M(I))(x))| = |((f \circ M(I))(x) - (g_f \circ M(I))(x))(f \circ M(I))(x) + (g_f \circ M(I))(x) - 2y| = |(f \circ M(I))(x) - (g_f \circ M(I))(x)| \cdot |(f \circ M(I))(x) + (g_f \circ M(I))(x) - 2y| \leq \alpha \cdot 2(b + \eta) = \varepsilon.
\]

Let \( x_1, \ldots, x_n \in \mathbb{R}^d \). The following result will allow us to relate the binary labeling number of the set \( \{M^Ix_1, \ldots, M^Ix_n\} \) to the binary labeling number of the set \( \{x_1(I), \ldots, x_n(I)\} \).

**Proposition 4.** Let \( A \) be a \( d \times k \) matrix with nonnegative entries. Let \( x_1, \ldots, x_n \in \mathbb{R}^d \). Then for \( m \geq 1 \),

\[
\Lambda(A^Tx_1, \ldots, A^Tx_n; m) \leq \Lambda(x_1, \ldots, x_n; m).
\]

**Proof.** Suppose \( x_i \preceq x_j \). Then also \( Ax_i \preceq Ax_j \). Therefore any labeling that is feasible for the points \( \{A^Tx_1, \ldots, A^Tx_n\} \) is also feasible for the points \( \{x_1, \ldots, x_n\} \).
We will now analyze the expected labeling number of a set of random variables \( \{W_1, \ldots, W_n\} \). The concept of an integer partition is required for the labeling number analysis.

**Definition 3 (Integer Partition)** (as stated in [5]). An integer partition of dimension \( (k - 1) \) with values in \( \{0, 1, \ldots, t\} \) is a collection of values \( A_{i_1, i_2, \ldots, i_{k-1}} \in \{0, 1, \ldots, t\} \) where \( i_l \in \{1, \ldots, m\} \) and \( A_{i_1, i_2, \ldots, i_{k-1}} \leq A_{j_1, j_2, \ldots, j_{k-1}} \) whenever \( i_l \leq j_l \) for all \( l \in \{1, \ldots, k-1\} \). The set of integer partitions of dimension \( (k - 1) \) with values in \( \{0, 1, \ldots, t\} \) is denoted by \( P\left(\left[t\right]^{k-1}\right) \).

For an illustration of a partition with \( k = 2 \), see Figure 3 in [5]. The following result provides a bound on the expected labeling number of a set of points with certain distribution assumptions. It will be used to bound the expected labeling number of the set \( \{X_1, \ldots, X_n\} \).

**Lemma 7.** Let \( m \in \mathbb{N} \) and \( B > 0 \). Let \( W \in \mathbb{R}^d \) be a random variable with support contained in the set \( [-B, B]^d \). Suppose that the density \( f_W(w) \) is upper-bounded by \( D \). Let \( W_1, \ldots, W_n \) be independent samples with distribution \( f_W \). Then

\[
\mathbb{E}[\Lambda(W_1, \ldots, W_n; m)] \leq \exp\left[\left(2 \log(2)(m - 1) + D^2 m + 2d - 1 B^d\right) n^{\frac{d-1}{m}}\right].
\]

**Proof.** Note that since the labeling number is translation-invariant, the same result applies to a random variable \( W \) with support contained in \( [0, 2B]^d \). [5] considered the case \( B = \frac{1}{2} \) and \( D = 1 \). We now adapt their proof. By Proposition 5 in [5],

\[
\Lambda(w_1, \ldots, w_n; m) \leq (\Lambda(w_1, \ldots, w_n; 2))^{m-1}
\]

for any \( w_1, \ldots, w_n \in \mathbb{R}^d \). For clarity of notation, write \( \Lambda(w_1, \ldots, w_n) \) in place of \( \Lambda(w_1, \ldots, w_n; 2) \). We therefore have

\[
\mathbb{E}[\Lambda(W_1, \ldots, W_n; m)] \leq \mathbb{E}\left[(\Lambda(W_1, \ldots, W_n))^{m-1}\right].
\]

Let \( t \in \mathbb{N} \). When \( B = \frac{1}{2} \) and \( D = 1 \), we have by Lemma 5 of [5]

\[
\mathbb{E}\left[(\Lambda(W_1, \ldots, W_n))^{m-1}\right] \leq |P([t]^d)|^{m-1} \mathbb{E}\left[2^{(m-1)N}\right],
\]

where \( N \sim \text{Binom}\left(n, \frac{t^d - (t-1)^d}{t^d}\right) \). The value \( \frac{t^d - (t-1)^d}{t^d} \) is the probability that a uniform random variable in \( [0, 1]^d \) falls in one of the \( t^d \) cubes out of \( t^d \) cubes that partition \( [0, 1]^d \). To adapt the proof, we instead partition \( [-B, B]^d \) into \( t^d \) cubes. Each cube therefore has volume \( \left(\frac{2B}{t}\right)^d \), and density upper-bounded by \( D \left(\frac{2B}{t}\right)^d \). We conclude that

\[
\mathbb{E}\left[(\Lambda(W_1, \ldots, W_n))^{m-1}\right] \leq |P([t]^d)|^{m-1} \mathbb{E}\left[2^{(m-1)N'}\right],
\]

where \( N' \sim \text{Binom}\left(n, D \left(\frac{2B}{t}\right)^d \left(t^d - (t-1)^d\right)\right) \). Let \( p = D \left(\frac{2B}{t}\right)^d \left(t^d - (t-1)^d\right) \).

[12] showed that

\[
|P([t]^d)| \leq \left(\frac{2t}{t}\right)^{t^d - 2}.
\]

We have

\[
\mathbb{E}\left[2^{(m-1)N'}\right] = \mathbb{E}\left[e^{\log(2)(m-1)N'}\right] = M_{N'}(\log(2)(m-1)),
\]

where \( M_{N'} \) is the moment-generating function of the random variable \( N' \). It holds that

\[
M_{N'}(\theta) = (1 - p + pe^\theta)^n.
\]

Substituting into (14),

\[
\mathbb{E}\left[(\Lambda(W_1, \ldots, W_n))^{m-1}\right] \leq \left(\frac{2t}{t}\right)^{t^d - 2} \left(1 - p + pe^{\log(2)(m-1)}\right)^n = \left(\frac{2t}{t}\right)^{(m-1)t^{d-2}} \left(1 - p + p2^{(m-1)}\right)^n.
\]
Using the fact that $\log(1 + x) \leq x$, we have
\[
E \left[ \left( \Lambda(W_1, \ldots, W_n) \right)^{m-1} \right] \leq \exp \left[ 2 \log(2)(m-1)t^{d-1} + np2^{m-1} \right] = \exp \left[ 2 \log(2)(m-1)t^{d-1} + D2^{m-1} \left( \frac{2B}{l} \right)^d \left( t^d - (t - 1)^d \right) n \right].
\]

By the Binomial Theorem,
\[
t^d - (t - 1)^d = t^d - \sum_{i=0}^{d} \binom{d}{i} t^{d-i}(-1)^i
\]
\[
= \sum_{i=1}^{d} \binom{d}{i} t^{d-i}(-1)^i + 1
\]
\[
\leq \sum_{i=1}^{d} \binom{k}{i} \max_{i \in \{1, \ldots, d\}} t^{d-i}(-1)^i + 1
\]
\[
= (2^d - 1)t^{d-1}.
\]

Substituting,
\[
E \left[ \left( \Lambda(W_1, \ldots, W_n) \right)^{m-1} \right] \leq \exp \left[ 2 \log(2)(m-1)t^{d-1} + D2^{m-1} \left( \frac{2B}{l} \right)^d \left( 2^d - 1 \right)t^{d-1}n \right]
\]
\[
= \exp \left[ 2 \log(2)(m-1)t^{d-1} + D2^{m-1} \left( \frac{2B}{l} \right)^d \left( 2^d - 1 \right)t^{d-1}n \right]
\]
\[
\leq \exp \left[ 2 \log(2)(m-1)t^{d-1} + D2^{m+2d-1}B^d t^{d-1}n \right].
\]

Let $t = n^{\frac{1}{d}}$. Substituting,
\[
E \left[ \left( \Lambda(W_1, \ldots, W_n) \right)^{m-1} \right] \leq \exp \left[ 2 \log(2)(m-1)n^{\frac{d-1}{d}} + D2^{m+2d-1}B^d n^{\frac{d}{d-1}} \right]
\]
\[
= \exp \left[ 2 \log(2)(m-1) + D2^{m+2d-1}B^d \right] n^{\frac{d}{d-1}}.
\]

We now prove Lemma 5. 

Proof of Lemma 5 Recall the definitions $G(T, R, I) = \{ f \circ (TR)^+(I) : f \in C(b) \}$ and $H(T, R, I) = \{ g \circ (TR)^+(I) : g \in C(b) \}$ for $R \in \mathcal{R}$. We have
\[
N_{G(T, R)}(\epsilon, n) \leq \sum_{R \in \mathcal{R}} \sum_{I \subset [d]: |I| = s^*} N_{G(T, R, I)}(\epsilon, n) \leq \binom{d}{s^*} \max_{R \in \mathcal{R}, I \subset [d]: |I| = s^*} N_{G(T, R, I)}(\epsilon, n).
\]

Consider an arbitrary $R \in \mathcal{R}$ and $I \subset [d]$ with $|I| = s^*$. By Proposition 3 the set $L_{H(T, R, I)}((x_1, y_1), \ldots, (x_n, y_n))$ is an $\epsilon$-net for the set $L_{G(T, R, I)}((x_1, y_1), \ldots, (x_n, y_n))$. Let $M = (TR)^+$. Observe that the cardinality of $L_{H(T, R, I)}((x_1, y_1), \ldots, (x_n, y_n))$ is upper-bounded by the labeling number of the set $\{(M(I))^T x_1, \ldots, (M(I))^T x_n\}$ with $\left( \left\lceil \frac{b}{\alpha} \right\rceil - 1 \right)$ labels. Recall the definition of $\overline{\pi}(I)$, and similarly let $\overline{\pi}(I)$ be the matrix formed from the rows of $M$ that are indexed by the set $I$. For $x \in \mathbb{R}^d$, it holds that
\[
(M(I))^T x = (\overline{\pi}(I))^T (\overline{\pi}(I)).
\]

Note that $\overline{\pi}(I)$ is an $s^*$-dimensional vector. By Proposition 4
\[
\Lambda \left( (M(I))^T x_1, \ldots, (M(I))^T x_n; \left( \left\lfloor \frac{b}{\alpha} \right\rfloor - 1 \right) \right) \leq \Lambda \left( \overline{x}_1(I), \ldots, \overline{x}_n(I); \left( \left\lfloor \frac{b}{\alpha} \right\rfloor - 1 \right) \right).
\]
Therefore, the value of $N_G(T, R, I) (\epsilon, n)$ is upper-bounded by the expected labeling number of the set \{$X_1(I), \ldots, X_n(I)$\} with \(\left(\frac{b}{\alpha} \right) - 1\) labels. Applying Lemma \[\text{setting } d = s^*, m = \left(\frac{b}{\alpha} \right) - 1, B = C \text{ and } D = (p^*)^s\], we have

$$N_G(T, R, I) (\epsilon, n) \leq \exp \left[2 \log(2) \left(\frac{b}{\alpha} - 2 + 2\left\lceil \frac{b}{\alpha} \right\rceil + 2s^* (p^*)^s \right) n^{\frac{s^*}{s}} \right]$$

$$\leq \exp \left[\frac{2 \log(2) b}{\alpha} + 2 \frac{b}{\alpha} (4p^*)^s \right] n^{\frac{s^*}{s}}.$$  

We conclude that

$$N_G(T, R) (\epsilon, n) \leq \left\lfloor \frac{d}{s^*} \right\rfloor |R| \exp \left[\frac{2 \log(2) b}{\alpha} + 2 \frac{b}{\alpha} (4p^*)^s \right] n^{\frac{s^*}{s}}.$$  

Recalling that $|R| = N_0^b$, we conclude.

**Remark 2.** In the proof of Lemma \[\text{we have taken advantage of the fact that } s^* \text{ is a constant. Another approach would be to analyze the labeling number directly in } \mathbb{R}^k, \text{ since } k \text{ is also a constant. However, there is a technical hurdle to overcome. The grid approach in Lemma \[\text{works well when the distribution of the random variable is not too concentrated. Without good control over the induced distribution of } M(I)^T X, \text{ it would be difficult to carry out a similar argument.}]

**Proof of Theorem \[** We show that Algorithm \[achieves the desired statistical guarantee, using the bound in Theorem \[We choose

$$N_0 = \left\lfloor \frac{\log(\frac{\epsilon}{\delta})}{\log \left(1 - \frac{s^*}{s} \frac{1}{\sqrt{\theta}} \right)} \right\rfloor \left\lfloor \frac{|S_k^{k-1} \cap B \left(e_1, \frac{\delta}{\sqrt{\theta}} \right)|}{|S_k^{k-1}|} \right\rfloor.$$  

The choice of $N_0$ leads to

$$k \left(1 - \frac{s^*}{s} \frac{1}{\sqrt{\theta}} \right) \leq \epsilon.$$

Similarly, \(\frac{1}{\sqrt{\theta}} \leq \frac{1}{\sqrt{s^*}}\) by the assumption on $d$.

Before bounding the last term, we need to control the value of $\epsilon_0$. Observe that the function $\varphi(\epsilon_1, \epsilon_2, C)$ is decreasing in both arguments. Therefore, by setting

$$\frac{1}{\rho^0} \leq \frac{1}{\sqrt{2s^*}} \lambda \leq \delta,$$

we ensure that $\epsilon_0 \geq \frac{1}{\sqrt{2}}$. Substituting in the value of $\lambda$ and solving,

$$\frac{1}{\rho^0} 4\sqrt{2} s^* \cdot 10 \sqrt{\theta} \frac{\log(d)}{n} \leq \delta$$

$$n \geq \frac{3200 s^2 \theta \log(d)}{\rho^0 \delta^2}.$$  

We now bound the last term:

$$\left\lfloor \frac{d}{s^*} \right\rfloor N_0 \exp \left[\frac{2 \log(2) b}{\alpha} + 2 \frac{b}{\alpha} (4p^*)^s \right] n^{\frac{s^*}{s}} - \frac{\epsilon_0 n}{2b^2} \right\rfloor \leq \exp \left[\log(4) + s^* \log(d) + \left(\frac{2 \log(2) b}{\alpha} + 2 \frac{b}{\alpha} (4p^*)^s \right) n^{\frac{s^*}{s}} - \frac{\epsilon_0 n}{2b^2} \right] \cdot \left(15\right).$$

Recall that $\alpha = \frac{1}{\sqrt{2}} \epsilon_0 (b + \eta)^{-1}$. For $n \geq 3200 s^2 \theta \log(d) \rho_0^{-2} \delta^{-2}$, we have $\epsilon_0 \geq \frac{1}{128} \epsilon_0 (b + \eta)^{-1}$. We see that there exists $t = t(N_0, C, b, s^*, p^*, k, \eta)$ such that if $n \geq \max \left\{t, 3200 s^2 \theta \log(d) \rho_0^{-2} \delta^{-2}\right\}$, then

$$\log(4) + \left(\frac{2 \log(2) b}{\alpha} + 2 \frac{b}{\alpha} (4p^*)^s \right) n^{\frac{s^*}{s}} - \frac{\epsilon_0 n}{2b^2} \leq \frac{\epsilon_0 n}{2b^2}.$$
For such \( n \), we then bound (15) by \( \exp \left[ s^* \log(d) - \frac{c_0 n}{2 \log^2 d} \right] \). Setting this quantity to be less than \( \frac{\epsilon}{3} \), we obtain

\[
\exp \left[ s^* \log(d) - \frac{c_0 n}{2 \log^2 d} \right] \leq \frac{\epsilon}{3}
\]

\[
s^* \log(d) - \frac{c_0 n}{2 \log^2 d} \leq \log \left( \frac{\epsilon}{3} \right)
\]

\[
n \geq \frac{2^{10} b^2}{c_0} \left( s^* \log(d) + \log \left( \frac{3}{\epsilon} \right) \right).
\]

Taking \( n_0 = \max \left\{ 2^{12} b^2 \epsilon^{-2} \left( s^* \log(d) + \log \left( \frac{3}{\epsilon} \right) \right), 3200 s^* 2^d \log(d) \rho_0^{-2} \delta^{-2}, t \right\} \) completes the proof. \( \square \)

8 Producing a Lipschitz estimator

We now modify Algorithm 2 to ensure the Lipschitz property, in addition to the coordinate-wise monotone property. When we only needed to ensure monotonicity, the interpolation step was straightforward; interpolation was possible as long as the points themselves satisfied the monotonicity property. The situation is slightly more complicated for Lipschitz functions. Algorithm 4 ensures that the estimated points are interpolable with respect to the class \( \mathcal{L}_1(b) \), as defined below.

**Definition 4.** We say that a collection of points \((x_i, y_i)_{i=1}^n \in \mathbb{R}^k \times \mathbb{R} \) is interpolable with respect to a function class \( \mathcal{F} \) if there exists \( f \in \mathcal{F} \) such that \( f(x_i) = y_i \) for each \( i \in [n] \).

Algorithm 4 below finds the optimal index set and function values on a given set of points, compatible with interpolability. Binary variables \( v_l \) determine the index set \( I \). The variables \( F_i \) represent the estimated function values at data points \( X_i \). Auxiliary variables \( z_{ij} \) and \( w_{ijp} \) are used to model the monotonicity and Lipschitz constraints.

**Algorithm 4** Integer Programming Sparse Matrix Isotonic Regression (Lipschitz)

**Input:** Values \((X_i, Y_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R} \), sparsity level \( s \), \( M \geq 0 \in \mathbb{R}^{d \times k} \), \( C > 0 \), \( b > 0 \)

**Output:** An index set \( I \subset [d] \) satisfying \( |I| = s \); values \( F_1, F_2, \ldots, F_n \in [0, b] \) such that the points \((M(I)^T X_i, F_i)_{i=1}^n \) are interpolable by a coordinate-wise monotone 1-Lipschitz function.

1. Let \( B = 2C \sum_{l=1}^d \sum_{p=1}^k M_{lp} \).
2. Solve the following optimization problem.

\[
\min_{v, F, z} \sum_{i=1}^n (Y_i - F_i)^2 \quad (16)
\]

s.t.

\[
\sum_{l=1}^d v_l = s \quad (17)
\]

\[
bz_{ij} \geq F_i - F_j \quad \forall i, j \in [n] \quad (18)
\]

\[
(F_i - F_j)^2 \leq \sum_{p=1}^k w_{ijp} \left( \sum_{l=1}^d v_l M_{lp}(X_{il} - X_{jl}) \right)^2 + (1 - z_{ij})b^2 \quad \forall i, j \in [n] \quad (19)
\]

\[
- B(1 - w_{ijp}) \leq \sum_{l=1}^d v_l M_{lp}(X_{il} - X_{jl}) \leq Bw_{ijp} \quad \forall i, j \in [n], p \in [k] \quad (20)
\]

\[
z_{ij} \in \{0, 1\} \quad \forall i, j \in [n]
\]

\[
v_l \in \{0, 1\} \quad \forall l \in [d]
\]

\[
w_{ijp} \in \{0, 1\} \quad \forall i, j \in [n], p \in [k]
\]

\[
F_i \in [0, b] \quad \forall i \in [n]
\]
Remark 3. Note that Constraints \((19)\) contain products of three binary variables. We may encode arbitrary products of binary variables using linear constraints, as follows. Suppose \(x\) and \(y\) are binary variables, and we wish to encode \(z = xy\). This is equivalent to the constraints \(x + y - 1 \leq z \leq \frac{1}{2}(x + y)\) and \(z \in \{0, 1\}\). Longer products may be encoded recursively.

We apply the construction of [11] to find a coordinate-wise monotone, 1-Lipschitz interpolation.

Proposition 5. Suppose the points \((x_i, y_i)_{i=1}^n \in \mathbb{R}^k \times [0, b]\) satisfy
\[
y_i - y_j \leq \| (x_i - x_j)^+ \|_2 \tag{21}
\]
for each pair \((i, j) \in [n]^2\). Let \(\hat{g}(x) = \max_i \{y_i - \| (x_i - x)^+ \|_2\}\), and let \(\hat{f}(x) = \max \{\hat{g}(x), 0\}\). Then \(\hat{f} \in L_1(b)\). Furthermore, \(\hat{f}\) interpolates the points; i.e. \(\hat{f}(x_i) = y_i\) for each \(i \in [n]\).

The proof follows from [11]. We therefore obtain the following approach for interpolation.

Algorithm 5 Monotone Lipschitz Interpolation

**Input:** Points \((x_i, y_i)_{i=1}^n \in \mathbb{R}^k \times [0, b]\) satisfying (21) for each \(i, j\).

**Output:** An estimated function \(\hat{f} \in L_1(b)\) that interpolates the points.

1. Let \(\hat{g}(x) = \max_i \{y_i - \| (x_i - x)^+ \|_2\}\). Return \(\hat{f}(x) = \max \{\hat{g}(x), 0\}\).

Algorithm 6

**Input:** Values \((X_i, Y_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}\), sparsity level \(s, M \geq 0 \in \mathbb{R}^{d \times k}\), \(C > 0, b > 0\)

**Output:** \(I_n \in [d] : |I_n| = s^*\) and \(f_n \in L_1(b)\)

1. Apply Algorithm 5 to input \((X_i, Y_i)_{i=1}^n, s^*, M, C, b\) obtaining the index set \(I\) and values \(F_1, \ldots, F_n\).
2. Apply Algorithm 5 to input \((M(I)X_{n+1}, F_i)_{i=1}^n\) obtaining the function \(\hat{f}\).
3. Return \((I, \hat{f})\).

Proposition 6. Suppose \(X_i \in [-C, C]^d\) for \(i \in [n]\). On input \((X_i, Y_i)_{i=1}^n, s, M, C, b\), Algorithm 6 finds a function \(\hat{f}_n \in L_1(b)\) and index set \(I_n\) that minimize the empirical loss \(\sum_{i=1}^n L(X_i, Y_i, f \circ M(I))\), over functions \(f \in L_1(b)\) and index sets \(I\) with cardinality \(s\).

We now modify Algorithm 5 to include the Lipschitz assumption.

Algorithm 7 MMI Regression (Lipschitz)

**Input:** \(N_0 \in \mathbb{N}\), values \((X_1, Y_1), \ldots, (X_N, Y_N)\), \(C > 0, b > 0, \tau > 0\), and \(\lambda > 0\)

**Output:** \(f_n \in L_1(b), Q_n \in \mathcal{C}_{d,k}, \mathcal{R}_n \in \mathcal{M}_{k,h}(r)\), and \(I_n \in [d] : |I_n| = s^*\)

1. Construct a random near-net \(\mathcal{R}(N_0)\).
2. Produce an estimate \(Q_n\) using Algorithm 1 applied to \((X_i, Y_i)_{i=1}^n, \tau, \text{and } \lambda\).
3. for each \(R \in \mathcal{R}\) do
   4. Let \(M = (Q_n R)^+\). Apply Algorithm 6 to input \((X_i, Y_i)_{i=1}^{n+1} \in \mathbb{R}^d \times \mathbb{R}, s^*, M, C, \text{and } b\) obtaining the index set \(I_R\) and function \(f_R\).
4. end for
5. Return the tuple \((f_R, Q_n, R, I_R)\) with the smallest empirical loss.

The proof of Theorem 3 carries through exactly for Algorithm 7 since \(L_1(b) \subset \mathcal{C}(b)\). We note that a tighter analysis of the estimation error incurred by using Algorithm 7 would take advantage of the Lipschitz property of the estimated function.

### 8.1 Proofs

In order to prove Proposition 5, we need to know when a collection of points is interpolable by a coordinate-wise monotone and 1-Lipschitz function. The following result provides a necessary and sufficient condition for interpolability.
We now show how the constraints encode this condition. First suppose \( \hat{f} \) interpolates the data also. Clearly \( \hat{g}(x) \geq 0 \) for all \( x \). Since \( y_i \) is free to equal \( x \), it holds that

\[
\hat{f}(x) \leq \max \{ b - \| (x_i - x) \|_2 \} \leq b \implies \hat{g}(x) \leq b.
\]

**Proof of Proposition 7** By Proposition 7 the data admits an interpolation by a coordinate-wise monotone 1-Lipschitz function. Further, the function \( \hat{f} \) interpolates the data, and is coordinate-wise monotone and 1-Lipschitz. Since \( y_i = 0 \) for all \( i \), the function \( \hat{g} \) interpolates the data also. The zero function is 1-Lipschitz and coordinate-wise monotone. Therefore, \( \hat{g} \), which is the pointwise maximum of two 1-Lipschitz and coordinate-wise monotone functions, is itself 1-Lipschitz and coordinate-wise monotone.

It remains to show that \( 0 \leq \hat{g}(x) \leq b \) for all \( x \). Clearly \( \hat{g}(x) \geq 0 \) for all \( x \). Since \( y_i \leq b \) for all \( i \), it holds that

\[
\hat{f}(x) \leq \max \{ b - \| (x_i - x) \|_2 \} \leq b \implies \hat{g}(x) \leq b.
\]

**Proof of Proposition 6** First we show that Algorithm 4 finds an index set \( I \). Constraint (17) ensures that exactly \( s \) of the \( v_i \) variables are set to 1, so that \( |I| = s \). Given this index set, we will show that the points \( (M(I)^T X_i, F_i)_{i=1}^n \) are interpolable by a coordinate-wise monotone 1-Lipschitz function. By Proposition 7 this is equivalent to

\[
F_i - F_j \leq \| (M(I)^T X_i - M(I)^T X_j) \|_2,
\]

for each \( i, j \). First observe that this is equivalent to either (a) \( F_i \leq F_j \) or

\[
(b) \quad (F_i - F_j)^2 \leq \| (M(I)^T X_i - M(I)^T X_j) \|_2^2.
\]

We now show how the constraints encode this condition. First suppose \( F_i > F_j \). Then by Constraint (18), \( z_{ij} = 1 \), and

\[
F_i - F_j \leq \| (M(I)^T X_i - M(I)^T X_j) \|_2 \iff (F_i - F_j)^2 \leq \| (M(I)^T X_i - M(I)^T X_j) \|_2^2 \iff (F_i - F_j)^2 \leq \| (M(I)^T X_i - M(I)^T X_j) \|_2^2 + (1 - z_{ij})b^2.
\]

Next suppose \( F_i \leq F_j \). Then \( z_{ij} \) is free to equal 0, so that

\[
(F_i - F_j)^2 \leq \| (M(I)^T X_i - M(I)^T X_j) \|_2^2 + (1 - z_{ij})b^2.
\]

We conclude that the points \( (M(I)^T X_i, F_i)_{i=1}^n \) are interpolable by a coordinate-wise monotone 1-Lipschitz function if and only if

\[
(F_i - F_j)^2 \leq \| (M(I)^T X_i - M(I)^T X_j) \|_2^2 + (1 - z_{ij})b^2.
\]

for each pair \( (i, j) \), where \( z_{ij} = 1 \) if \( F_i > F_j \). Expanding the right hand side of (22),

\[
\| (M(I)^T X_i - M(I)^T X_j) \|_2^2 = \sum_{p=1}^k \left( \sum_{l=1}^d (M(I)^T)_{pl} (X_{il} - X_{jl}) \right)^{+,2} = \sum_{p=1}^k \left( \sum_{l=1}^d v_l M_{lp} (X_{il} - X_{jl}) \right)^{+,2}.
\]
Now, Constraints (20) ensure that
\[
\begin{align*}
    w_{ijp} &= \begin{cases} 
        1 & \text{if } \sum_{l=1}^{d} v_l M_{lp}(X_{il} - X_{jl}) > 0 \\
        0 & \text{if } \sum_{l=1}^{d} v_l M_{lp}(X_{il} - X_{jl}) < 0
    \end{cases}. 
\end{align*}
\]
Therefore,
\[
\| (M(I)^T X_i - M(I)^T X_j)^+ \|_2^2 = \sum_{p=1}^{k} w_{ijp} \left( \sum_{l=1}^{d} v_l M_{lp}(X_{il} - X_{jl}) \right)^2. 
\]
We conclude that the points \((M(I)^T X_i, F_i)_{i=1}^{n}\) are interpolable by a coordinate-wise monotone 1-Lipschitz function if and only if
\[
(F_i - F_j)^2 \leq \sum_{p=1}^{k} w_{ijp} \left( \sum_{l=1}^{d} v_l M_{lp}(X_{il} - X_{jl}) \right)^2 + (1 - z_{ij}) b^2, 
\]
for each \((i, j)\), which is exactly Constraints (19).

The objective (16) minimizes the loss on the samples. Finally, the interpolation step produces a coordinate-wise monotone Lipschitz function with range \([0, b]\), by Proposition 5. \(\square\)