Learning Multi-index Models

Giannis Iakovidis

TTIC, Summer 2025

This Talk Is Based On

- Robust Learning of Multi-index Models via Iterative Subspace Approximation [DIKZ25]
  I. Diakonikolas, G. Iakovidis, D. Kane, N. Zarifis

- Algorithms and SQ Lower Bounds for Robustly Learning Real-valued Multi-index Models [DIKR25]
  I. Diakonikolas, G. Iakovidis, D. Kane, R. Lisheng

Multi-Index Models

Definition (Multi-Index Models (MIMs))

A class of function $\mathcal{F} \subseteq \{f : \mathbb{R}^d \to \mathcal{Y}\}$ is called a class of MIMs of dimension K, if for every $f \in \mathcal{F}$ there exists a subspace $W \subseteq \mathbb{R}^d$, of dimension at most K such that $f(x) = f(x^W)$.

- Essentially each function depends on the projection onto a low dimensional subspace W. Can be written as $f(Wx)$.
- We assume that $K \ll d$.
- We assume that the label space is finite, $|\mathcal{Y}| < \infty$.
- Many well-studied function classes, such as neural networks, multiclass linear classifiers, intersections of halfspaces are MIM classes.

Example MIMs

$$\prod_{i \in [K]} \mathbb{1}\left(w^{(i)} \cdot x + t_i \geq 0\right)$$

$$\operatorname{argmax}_{i \in [K]}\left(w^{(i)} \cdot x + t_i\right)$$

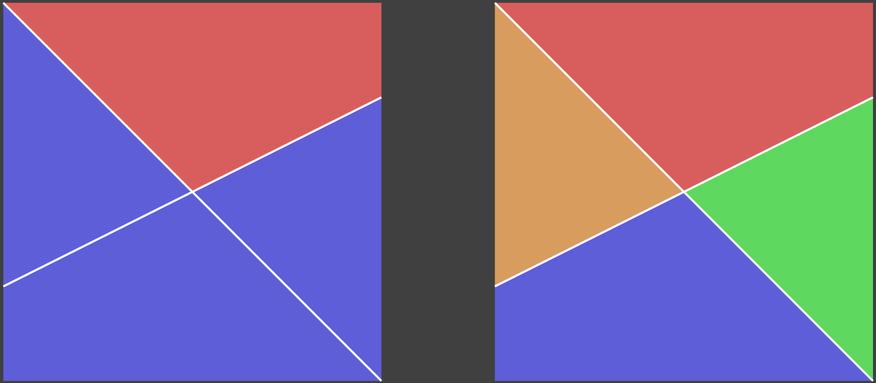


Figure 1: Intersection of halfspaces and Linear Multiclass Classifiers

Example MIMs

Figure 2: Homogeneous ReLU Network

Setting

- We will work in the agnostic label noise setting.
- We observe samples (x, y), where $x \sim D$ and y equals $f(x)$ except in an OPT fraction of the samples.
- Our goal is to find a function h such that $\mathbf{Pr}[y \neq h(x)]$ is small.

Setting

- We will work in the agnostic label noise setting.
- We observe samples (x, y), where $x \sim D$ and y equals $f(x)$ except in an OPT fraction of the samples.
- Our goal is to find a function h such that $\mathbf{Pr}[y \neq h(x)]$ is small.

However this is computationally hard!!
We need assumptions on the function f, the distribution D and a relaxed error guarantee.

Learning Goal

We will work in the agnostic label noise setting with distributional assumptions.

- Let D be a distribution over $\mathbb{R}^d \times \mathcal{Y}$ with $D_x = \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- Let $\mathcal{F}$ be a MIM class, e.g., multiclass linear classifiers.

Learning Goal

We will work in the agnostic label noise setting with distributional assumptions.

- Let D be a distribution over $\mathbb{R}^d \times \mathcal{Y}$ with $D_x = \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- Let $\mathcal{F}$ be a MIM class, e.g., multiclass linear classifiers.

Goal:

- Given N samples, $(x^{(i)}, y_i) \sim D$ for $N = \mathrm{poly}(d) \cdot g(\epsilon, K, \delta)$.
- Find an algorithm that runs in $\mathrm{poly}(N)$ time and returns a hypothesis h comparable with the best-in-class

$$\mathbf{Pr}_{(x,y)\sim D}[h(x) \neq y] \leq c(K, \mathrm{OPT}) + \epsilon \text{ w.p. } 1 - \delta ,$$

c is a small function of K and $\mathrm{OPT} = \inf_{f \in \mathcal{F}} \mathbf{Pr}_{(x,y)\sim D}[f(x) \neq y]$.

Main Result

Theorem (Informal Main Theorem)

There exists a dimension-efficient and robust algorithm for broad family of well-behaved MIMs. Moreover, there is a SQ lower bound demonstrating that this algorithm is optimal wrt the dependence on the dimension.

Multiclass Linear Classification

Theorem (Agnostically Learning $\mathcal{L}_{d,K}$)

There exists an algorithm that draws $N = d\, 2^{\mathrm{poly}(K/\epsilon)}$ i.i.d. labeled samples, runs in $\mathrm{poly}(N)$ time, and outputs a hypothesis h such that w.h.p. $\mathrm{err}_{0-1}^{D}(h) \le O(\mathrm{OPT}) + \epsilon$, where $\mathrm{OPT} = \inf_{f \in \mathcal{L}_{d,K}} \mathrm{err}_{0-1}^{D}(f)$.

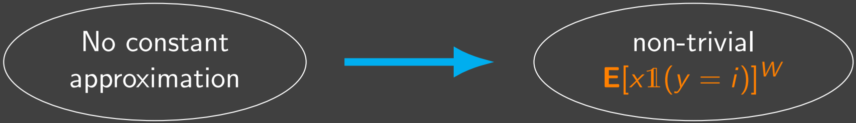
Intuitively, the algorithm will approximately recover the subspace W using moments, i.e., $\mathbf{E}[x\mathbb{1}(y = i)]$. Subsequently, it performs a brute-force search within the recovered subspace.

Finding a relevant direction

- If there exists a label i such that $\Pr[i \neq y] \leq C\mathrm{OPT} + \epsilon$ we could just output i.
- Otherwise, assuming that y is far from a constant, we have that each class has a substantial first moment that the adversary can not hide.

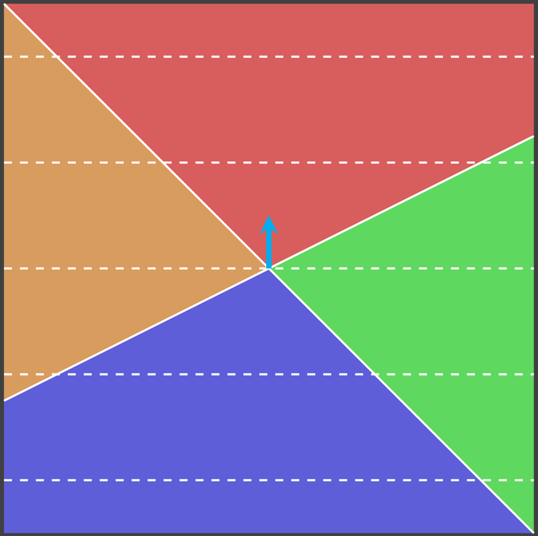
Finding a relevant direction

- If there exists a label i such that $\Pr[i \neq y] \leq C\mathrm{OPT} + \epsilon$ we could just output i.
- Otherwise, assuming that y is far from a constant, we have that each class has a substantial first moment that the adversary can not hide.

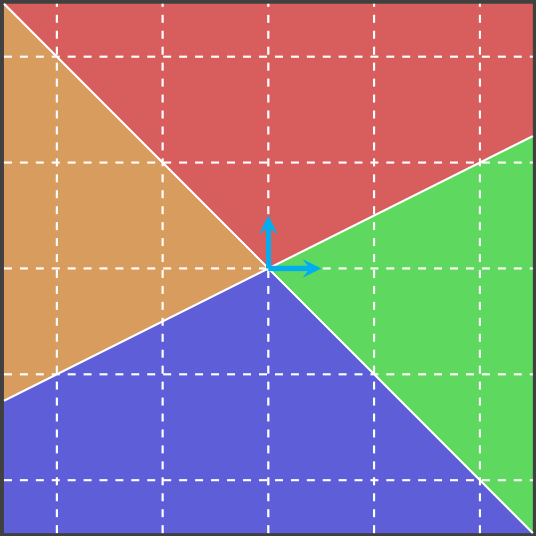


Hence we have recovered one relevant direction!!

Iterative Approximation



Iterative Approximation



Algorithm

1. Let $L_1 \leftarrow \varnothing$
2. for $t = 1 : T$
3. Form a partition S_t of $\mathrm{span}(L_t)$ into cubes.
4. Set $L_{t+1} \leftarrow L_t \cup \{\mathbf{E}[x\mathbb{1}(y = i) \mid x \in S]\}_{S \in S_t}$
5. Form a partition S_T of $\mathrm{span}(L_T)$ into cubes.
6. Return h a function that outputs the most frequent label for every cube.

Algorithm

1. Let $L_1 \leftarrow \varnothing$
2. for $t = 1 : T$
3. Form a partition S_t of $\mathrm{span}(L_t)$ into cubes.
4. Set $L_{t+1} \leftarrow L_t \cup \{\mathbf{E}[x\mathbb{1}(y = i) \mid x \in S]\}_{S \in S_t}$
5. Form a partition S_T of $\mathrm{span}(L_T)$ into cubes.
6. Return h a function that outputs the most frequent label for every cube.

If T is a sufficiently large polynomial of K and $1/\epsilon$ and you take enough samples to approximate the expectations accurately, then h achieves $O(\mathrm{OPT}) + \epsilon$ error.

Analysis

1. Let $L_1 \leftarrow \varnothing$
2. for $t = 1 : T$
3. Form a partition S_t of $\mathrm{span}(L_t)$ into cubes.
4. Set $L_{t+1} \leftarrow L_t \cup \{\mathbf{E}[x\mathbb{1}(y = i) \mid x \in S]\}_{S \in S_t}$
5. Form a partition S_T of $\mathrm{span}(L_T)$ into cubes.
6. Return h a function that outputs the most frequent label for every cube.

- To approximate $\mathbf{E}[x\mathbb{1}(y = i)]$ accurately you need d/ϵ^2 samples.
- If you have ϵ as the width of the cube and $\dim(\mathrm{span}(L_t)) = k_t$ then $|S_t| = \frac{1}{\epsilon^{k_t}}$. The number of cubes increases uncontrollably and so does the complexity.
- We need a filtering step!!

Filtering

- In fact we show that there is an ϵ fraction of cubes S with $w_i \cdot \mathbf{E}[x\mathbb{1}(y = j) \mid x \in S]$ non-trivial for some i and j.
- So there is an ϵ/K fraction of the cubes that have non-trivial moments for the same w_i.

Filtering

- In fact we show that there is an ϵ fraction of cubes S with $w_i \cdot \mathbf{E}[x\mathbb{1}(y=j) \mid x \in S]$ non-trivial for some i and j.
- So there is an ϵ/K fraction of the cubes that have non-trivial moments for the same w_i.
- Therefore the matrix

$$U = \sum_{S,i} u_{S,i} u_{S,i}^\top \mathbf{Pr}[x \in S] \quad u_{S,i} = \mathbf{E}[x\mathbb{1}(y=i) \mid x \in S]$$

  has a big quadratic form for some w_i.
- Consequently we can find a vector close to w in U's largest eigenvalues.
- This reduces the number of added vectors to $\mathrm{poly}(K/\epsilon)$.

General Algorithm?

What properties of the function class have we used?

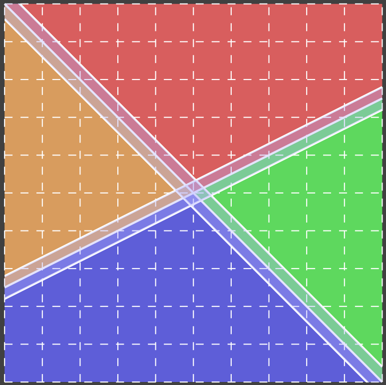
1. Let $L_1 \leftarrow \varnothing$
2. for $t = 1 : T$
3. Form a partition S_t of $\mathrm{span}(L_t)$ into cubes.
4. Set $L_{t+1} \leftarrow L_t \cup \text{Filter}\left(\{\mathbf{E}[x\mathbb{1}(y = i) \mid x \in S]\}_{S \in S_t}\right)$

   1. Existence of correlating moments
5. Form a partition S_T of $\mathrm{span}(L_T)$ into cubes.
6. Return h a function that outputs the most frequent label for every cube.

   2. Approximability of f from cubes

Approximability from cubes

- For real-valued functions Lipschitzness ($\|\nabla f(x)\| \leq L$) or more generally bounded total variation ($\mathbf{E}[\|\nabla f(x)\|] \leq L$) suffices.
- For discrete-valued functions under the gaussian the analogous measure is the Gaussian Surface Area.
- It is a measure of complexity of the decision boundary.



Well-Behaved MIMs

Definition (Well-Behaved K-MIM)

Let $f : \mathbb{R}^d \to \mathcal{Y}$ be a K-MIM. We say that f is (m, ζ, τ, Γ)-well-behaved if the following two conditions hold:

1. The Gaussian surface area of the decision region of f is at most Γ.
2. For every joint distribution (x, y) on $\mathbb{R}^d \times \mathcal{Y}$ satisfying

$$\Pr_{(x,y)}\big[f(x) \neq y\big] \;\leq\; \zeta$$

   and for every linear subspace $V \subseteq \mathbb{R}^d$, one of the following is true:

   1. $\Pr\big[f(x) \neq g(x^V)\big] \;\leq\; \tau$.
   2. With non-trivial probability over $x \in V$, conditioned on that point, the resulting conditional distribution of x has a non-vanishing moment of degree at most m.

Main Theorem

Theorem (General Algorithm)

There exists an agnostic learning algorithm for (m, ζ, τ, Γ)*-well-behaved MIMs that, where* $\zeta \geq \mathrm{OPT} + \epsilon$ *that, uses* $N = d^m 2^{\mathrm{poly}\left(\Gamma K |\mathcal{Y}|/\epsilon\right)}$ *samples, runs in time* $\mathrm{poly}(N)$*, and outputs a hypothesis* h *satisfying, with probability at least* $1 - \delta$*,*

$$\mathrm{err}_{0\text{–}1}^{D}(h) \leq \tau + \mathrm{OPT} + \epsilon.$$

Furthermore we prove:

- That $N = d^m \mathrm{poly}(\Gamma |\mathcal{Y}|/\epsilon)^K$ suffices when y depends only on W.
- A matching lower bound for classes of functions that do not satisfy the well-behaved MIM condition.

Real-Valued Concepts

Definition (Well-Behaved K-MIM)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a K-MIM. f is (m, ζ, τ, L, M)-well-behaved if the following two conditions hold:

1. $\mathbf{E}[f^2(x)] \le M, \mathbf{E}_{x \sim \mathcal{N}(0,I)}[\|\nabla f(x)\|^2] \le L$.
2. For every joint distribution (x, y) on $\mathbb{R}^d \times \mathcal{Y}$ satisfying

$$\Pr_{(x,y)}\left[(f(x) - y)^2\right] \le \zeta$$

   and for every linear subspace $V \subseteq \mathbb{R}^d$, one of the following is true:

   1. $\Pr\left[(f(x) - g(x^V))^2\right] \le \tau$.
   2. There exists a point $x \in V$ such that, conditioned on that point, the resulting conditional distribution of x has a non-vanishing moment of degree at most m.

Real-Valued Concepts

- Essentially these conditions allow you to bin the real-valued labels into intervals of non-trivial length and run the same algorithm.
- These conditions lead to the same characterization of efficient learnability of MIMs.
- The matching SQ lower bound is more challenging to prove since we can have very large chi-squared divergence with $\mathcal{N}(\mathbf{0}, \mathbf{I})$. But along with prior work [DKRS23] that focused on the unsupervised setting we developed tools that deal with this issue.

Results for Well-Studied Function Classes

By applications of the general theorem we have proven new guarantees for many well-studied function classes:

Function Class	**Runtime**	**Error**
Agnostic K-MLC	$\mathrm{poly}(d)\, 2^{\mathrm{poly}(K/\epsilon)}$	$O(\mathrm{OPT}) + \epsilon$
K-MLC with RCN	$\mathrm{poly}(d)\, (1/\epsilon)^{\mathrm{poly}(K)}$	$O(\mathrm{OPT}) + \epsilon$
Agnostic Intersections of K halfspaces	$\mathrm{poly}(d)\, 2^{\mathrm{poly}(K/\epsilon)}$	$K\, \tilde{O}(\mathrm{OPT}) + \epsilon$
Well-Behaved K-MIMs	$d^{O(m)}\, 2^{\mathrm{poly}(mK\Gamma/\epsilon)}$	$\tau + \mathrm{OPT} + \epsilon$
Positive Hom. & Lipschitz Functions	$\mathrm{poly}(d) 2^{\mathrm{poly}(KL/\epsilon)}$	ϵ

Thank you for your attention.
Are there any questions?

Bibliography I

[DIKR25] I. Diakonikolas, G. Iakovidis, D. M. Kane, and L. Ren. "Algorithms and SQ Lower Bounds for Robustly Learning Real-valued Multi-index Models". In: Arxiv. 2025.

[DIKZ25] I. Diakonikolas, G. Iakovidis, D. M. Kane, and N. Zarifis. "Robust Learning of Multi-index Models via Iterative Subspace Approximation". In: Arxiv. 2025.

[DKRS23] I. Diakonikolas, D. Kane, L. Ren, and Y. Sun. "SQ Lower Bounds for Non-Gaussian Component Analysis with Weaker Assumptions". In: Neurips. 2023.