Robust Learning of Multi-index Models via Iterative Subspace Approximation

Ilias Diakonikolas Giannis Iakovidis Daniel Kane Nikos Zarifis

FOCS 2025

Multi-Index Models

$x^W := \mathrm{Proj}_W x$

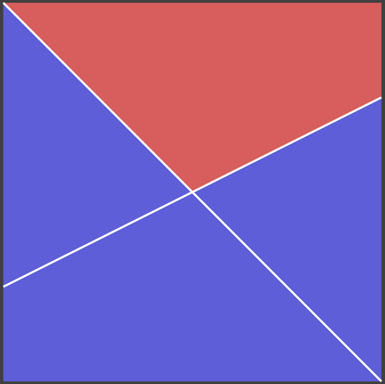
Definition (Multi-Index Models (MIMs))

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

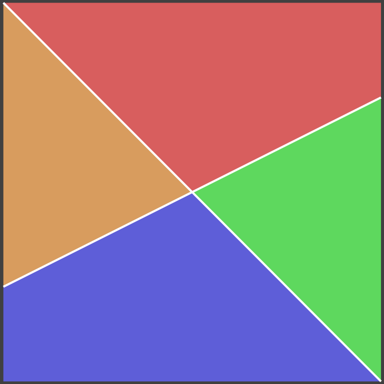
- Essentially each function depends on the projection onto a low dimensional subspace.
- We assume that $K \ll d$.
- Many well-studied function classes, such as neural networks, multiclass linear classifiers, intersections of halfspaces are MIM classes.
- We assume that the label space is finite, $|\mathcal{Y}|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

Example MIMs

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



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



Setting

- We will work in the agnostic label noise setting.
- We observe samples (x, y) from an unknown distribution D and we want to return a classifier h that has $0-1$ error comparable to $\mathrm{OPT} := \inf_{f \in \mathcal{F}} \Pr[f(x) \neq y]$

Setting

- We will work in the agnostic label noise setting.
- We observe samples (x, y) from an unknown distribution D and we want to return a classifier h that has $0-1$ error comparable to $\mathrm{OPT} := \inf_{f \in \mathcal{F}} \Pr[f(x) \neq y]$

We will assume that $D_x = \mathcal{N}(0, I)$.

Setting

- We will work in the agnostic label noise setting.
- We observe samples (x, y) from an unknown distribution D and we want to return a classifier h that has $0-1$ error comparable to $\mathrm{OPT} := \inf_{f \in \mathcal{F}} \Pr[f(x) \neq y]$

We will assume that $D_x = \mathcal{N}(0, I)$.
Getting $\mathrm{OPT} + \epsilon$ needs $d^{\mathrm{poly}(1/\epsilon)}$ complexity even for learning halfspaces.
[Diakonikolas-Kane-Pittas-Zarifis'21, Diakonikolas-Kane-Ren'23]

Setting

- We will work in the agnostic label noise setting.
- We observe samples (x, y) from an unknown distribution D and we want to return a classifier h that has $0-1$ error comparable to $\mathrm{OPT} := \inf_{f \in \mathcal{F}} \Pr[f(x) \neq y]$

We will assume that $D_x = \mathcal{N}(0, I)$.
Getting $\mathrm{OPT} + \epsilon$ needs $d^{\mathrm{poly}(1/\epsilon)}$ complexity even for learning halfspaces.
[Diakonikolas-Kane-Pittas-Zarifis'21, Diakonikolas-Kane-Ren'23]
We aim for a relaxed error guarantee that gets $\mathrm{poly}(d)$-dependence.

Learning Goal

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

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 $N = \mathrm{poly}(d) \cdot g(\epsilon, K)$.
- Find an algorithm that runs in $\mathrm{poly}(N)$ time and returns a hypothesis h comparable with the best-in-class

$$\Pr_{(x,y)\sim D}[h(x) \neq y] \leq c(K) f(\mathrm{OPT}) + \epsilon \,,$$

where c, f small functions of K and OPT.

Main Results

Theorem (Informal Main Theorem)

Dimension-efficient and robust algorithm for broad family of well-behaved MIMs. Matching Statistical Query lower bound.

Main Results

Theorem (Informal Main Theorem)

Dimension-efficient and robust algorithm for broad family of well-behaved MIMs. Matching Statistical Query lower bound.

Theorem (Agnostically Learning MLC)

There exists an algorithm that draws $N = d\,2^{\mathrm{poly}(K/\epsilon)}$ *samples, runs in* $\mathrm{poly}(N)$ *time, and returns* h *such that w.h.p.* $\mathrm{err}_{0-1}^{D}(h) \leq O(\mathrm{OPT}) + \epsilon$.

Main Results

Theorem (Informal Main Theorem)

Dimension-efficient and robust algorithm for broad family of well-behaved MIMs. Matching Statistical Query lower bound.

Theorem (Agnostically Learning MLC)

There exists an algorithm that draws $N = d\, 2^{\mathrm{poly}(K/\epsilon)}$ *samples, runs in* $\mathrm{poly}(N)$ *time, and returns* h *such that w.h.p.* $\mathrm{err}_{0-1}^{D}(h) \leq O(\mathrm{OPT}) + \epsilon$.

Theorem (Agnostically Learning Intersections of K-halfspaces)

There exists an algorithm that draws $N = d^2 2^{\mathrm{poly}(K/\epsilon)}$ *samples, runs in* $\mathrm{poly}(N)$ *time, and returns* h *such that w.h.p.* $\mathrm{err}_{0-1}^{D}(h) \leq K\,\widetilde{O}(\mathrm{OPT}) + \epsilon$.

Intersections of K-halfspaces

Theorem (Agnostically Learning Intersections of K-halfspaces)

There exists an algorithm that draws $N = d^2 2^{\mathrm{poly}(K/\epsilon)}$ samples, runs in $\mathrm{poly}(N)$ time, and returns h such that w.h.p. $\mathrm{err}_{0-1}^{D}(h) \leq K\widetilde{O}(\mathrm{OPT}) + \epsilon$.

[Diakonikolas-Kane-Stewart'18] gives error $\mathrm{poly}(K)\widetilde{O}(\mathrm{OPT}^{1/11}) + \epsilon$ with complexity $\mathrm{poly}(d)/\epsilon^{\mathrm{poly}(K)}$.

Multiclass Linear Classification (MLC)

Theorem (Agnostically Learning MLC)

There exists an algorithm that draws $N = d\,2^{\mathrm{poly}(K/\epsilon)}$ *samples, runs in* $\mathrm{poly}(N)$ *time, and returns* h *such that w.h.p.* $\mathrm{err}_{0-1}^{D}(h) \leq O(\mathrm{OPT}) + \epsilon$.

Multiclass Linear Classification (MLC)

Theorem (Agnostically Learning MLC)

There exists an algorithm that draws $N = d\,2^{\mathrm{poly}(K/\epsilon)}$ *samples, runs in* $\mathrm{poly}(N)$ *time, and returns* h *such that w.h.p.* $\mathrm{err}_{0-1}^{D}(h) \leq O(\mathrm{OPT}) + \epsilon$.

- Efficiently solvable in the distribution-free realizable setting using LP.
- For $K = 2$ and agnostic noise we have good understanding of the distribution specific setting.
  [Awasthi-Balcan-Long'17, Diakonikolas-Kane-Stewart'18, Diakonikolas-Kontonis-Tzamos-Zarifis'20]

Multiclass Linear Classification (MLC)

Theorem (Agnostically Learning MLC)

There exists an algorithm that draws $N = d\,2^{\mathrm{poly}(K/\epsilon)}$ samples, runs in $\mathrm{poly}(N)$ time, and returns h such that w.h.p. $\mathrm{err}_{0-1}^{D}(h) \leq O(\mathrm{OPT}) + \epsilon$.

- Efficiently solvable in the distribution-free realizable setting using LP.
- For $K = 2$ and agnostic noise we have good understanding of the distribution specific setting.
  [Awasthi-Balcan-Long'17, Diakonikolas-Kane-Stewart'18, Diakonikolas-Kontonis-Tzamos-Zarifis'20]
- For $K > 2$ and noise nothing was known algorithmically.

Standard Dimension-Reduction

Two-step procedure:

1. Find approximation V to the defining subspace W.
2. Use exhaustive search over V.

Standard Dimension-Reduction

Two-step procedure:

1. Find approximation V to the defining subspace W.
2. Use exhaustive search over V.

Dimension-reduction:

1. Estimate low-order moments of the level-sets of y

$$\mathop{\mathbf{E}}_{(x,y)}\left[p(x)\cdot \mathbb{1}(y=i)\right]$$

for all low-degree polynomials p and labels i.

2. Use moments to extract V.

Standard Dimension-Reduction

Two-step procedure:

1. Find approximation V to the defining subspace W.
2. Use exhaustive search over V.

Dimension-reduction:

1. Estimate low-order moments of the level-sets of y

$$\mathop{\mathbf{E}}_{(x,y)}\left[p(x)\cdot\mathbb{1}(y=i)\right]$$

   for all low-degree polynomials p and labels i.
2. Use moments to extract V.

We will consider conditional moments $\mathbf{E}[x\mathbb{1}(y=i) \mid x \in R]$ and an iterative method.

Finding a relevant direction

Lemma (Structural Result for Multiclass Linear Classification)

If f is not $\mathrm{COPT} + \epsilon$ close to being constant, then there exists a label $i \in [K]$ such that:

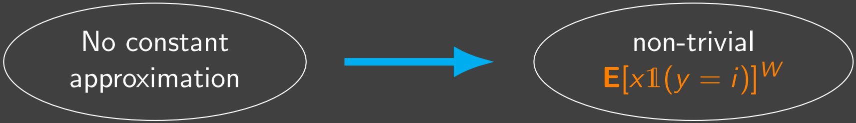
$$\left\| \mathbb{E}\left[x \mathbb{1}(y = i)\right]^{W} \right\| \geq \mathrm{poly}(\epsilon / K)$$

Finding a relevant direction

Lemma (Structural Result for Multiclass Linear Classification)

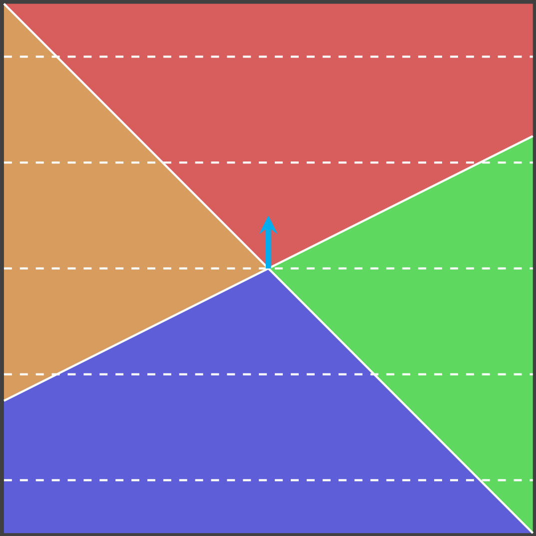
If f is not $\mathrm{COPT} + \epsilon$ close to being constant, then there exists a label $i \in [K]$ such that:

$$\left\| \mathbb{E}\left[x\mathbb{1}(y = i)\right]^W \right\| \geq \mathrm{poly}(\epsilon/K)$$

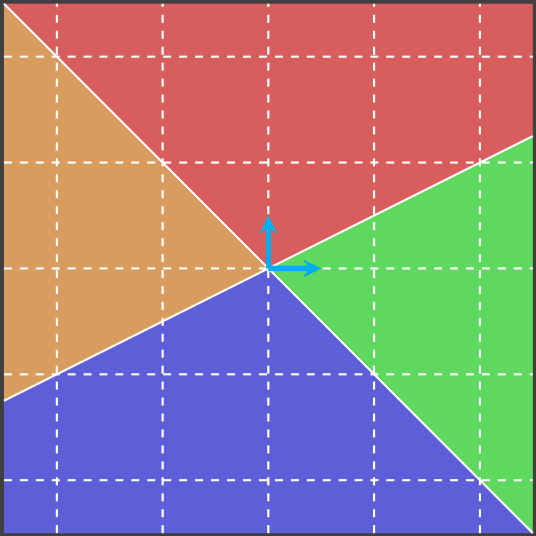


We have recovered one relevant direction.

Iterative Approximation



Iterative Approximation



Algorithm: First Attempt

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 the best piecewise constant approximation h of y over cubes.

Algorithm: First Attempt

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 the best piecewise constant approximation h of y over cubes.

Proposition

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 the best piecewise constant approximation h of y over cubes.

- For cubes of width ϵ and $\dim(\mathrm{span}(L_t)) = k_t$, $|S_t| = \frac{1}{\epsilon^{k_t}}$.
  The number of cubes increases exponentially at every iteration.
- Idea: Spectral pruning step that adds only $\mathrm{poly}(K/\epsilon)$ directions.

General 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 $L_{t+1} \leftarrow L_t \cup \text{Prune}(\{\mathbf{E}[x\mathbb{1}(y = i) \mid x \in S]\}_{S \in S_t})$
5. Form a partition S_T of $\mathrm{span}(L_T)$ into cubes.
6. Return the best piecewise constant approximation h of y over cubes.

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 $L_{t+1} \leftarrow L_t \cup \text{Prune}(\{\mathbf{E}[x\mathbb{1}(y = i) \mid x \in S]\}_{S \in S_t})$
5. Form a partition S_T of $\mathrm{span}(L_T)$ into cubes.
6. Return the best piecewise constant approximation h of y over cubes.

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 $L_{t+1} \leftarrow L_t \cup \text{Prune}(\{\mathbf{E}[x\mathbb{1}(y = i) \mid x \in S]\}_{S \in S_t})$

   1. Existence of correlating moments
5. Form a partition S_T of $\mathrm{span}(L_T)$ into cubes.
6. Return the best piecewise constant approximation h of y over cubes.

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 $L_{t+1} \leftarrow L_t \cup \mathsf{Prune}(\{\mathbf{E}[x\mathbb{1}(y = i) \mid x \in S]\}_{S \in \mathcal{S}_t})$
   1. Existence of correlating moments
5. Form a partition S_T of $\mathrm{span}(L_T)$ into cubes.
6. Return the best piecewise constant approximation h of y over cubes.
   2. Existence of efficient piecewise-constant approximation

Well-Behaved MIMs

We assume bounded Gaussian Surface Area [Klivans-O'Donnell-Servedio'08].

Definition (Well-Behaved K-MIM)

A K-MIM $f : \mathbb{R}^d \to \mathcal{Y}$ is (m, ζ, τ)-well-behaved if for any distribution (x, y) such that $\Pr_{(x,y)}[f(x) \neq y] \leq \zeta$ and any subspace $V \subseteq \mathbb{R}^d$ we have:

1. either $\Pr[f(x) \neq g(x^V)] \leq \tau$
2. or there exists a point in V such that if we condition on that point, there exists a non-vanishing moment of degree at most m.

Well-Behaved MIMs

We assume bounded Gaussian Surface Area [Klivans-O'Donnell-Servedio'08].

Definition (Well-Behaved K-MIM)

A K-MIM $f : \mathbb{R}^d \to \mathcal{Y}$ is (m, ζ, τ)-well-behaved if for any distribution (x, y) such that $\Pr_{(x,y)}[f(x) \neq y] \leq \zeta$ and any subspace $V \subseteq \mathbb{R}^d$ we have:

1. either $\Pr[f(x) \neq g(x^V)] \leq \tau$
2. or there exists a point in V such that if we condition on that point, there exists a non-vanishing moment of degree at most m.

We have proven:

- Multiclass linear classifiers are $(1, \mathrm{OPT}, O(\mathrm{OPT}))$-well-behaved.
- Intersections of K halfspace are $(2, \mathrm{OPT}, K\widetilde{O}(\mathrm{OPT}))$-well-behaved.

Characterization of Efficient Learnability of MIMs

Let $\mathcal{F}$ be a K-MIM family, and let $\zeta, \tau > 0$.

$$m^* \;:=\; \min\big\{ m \in \mathbb{Z}_{\geq 0} : \text{ every } f \in \mathcal{F} \text{ is } (m, \zeta, \tau)\text{-well-behaved} \big\}$$

Characterization of Efficient Learnability of MIMs

Let $\mathcal{F}$ be a K-MIM family, and let $\zeta, \tau > 0$.

$$m^* \; := \; \min\left\{ m \in \mathbb{Z}_{\geq 0} : \; \text{every } f \in \mathcal{F} \text{ is } (m, \zeta, \tau)\text{-well-behaved} \right\}$$

SQ Characterization

There is an efficient SQ algorithm that learns any ζ-noisy function from $\mathcal{F}$ to error $\tau + O(\zeta) + \epsilon$ using

$$N \; = \; d^{\,m^*} \, 2^{\mathrm{poly}(K|\mathcal{Y}|/\epsilon)}$$

samples and time $\mathrm{poly}(N)$. Moreover, no efficient SQ algorithm can achieve error $\tau - O(\zeta)$ with resources $d^{o(m^*)}$.

Characterization of Efficient Learnability of MIMs

Let $\mathcal{F}$ be a K-MIM family, and let $\zeta, \tau > 0$.

$$m^* \;:=\; \min\left\{ m \in \mathbb{Z}_{\geq 0} : \text{ every } f \in \mathcal{F} \text{ is } (m, \zeta, \tau)\text{-well-behaved} \right\}$$

SQ Characterization

There is an efficient SQ algorithm that learns any ζ-noisy function from $\mathcal{F}$ to error $\tau + O(\zeta) + \epsilon$ using

$$N \;=\; d^{\,m^*}\, 2^{\mathrm{poly}(K|\mathcal{Y}|/\epsilon)}$$

samples and time $\mathrm{poly}(N)$. Moreover, no efficient SQ algorithm can achieve error $\tau - O(\zeta)$ with resources $d^{o(m^*)}$.

Corollary

poly(d) learnability occurs if and only if $m^* = O(1)$.

Characterization of Efficient Learnability of MIMs

Let $\mathcal{F}$ be a K-MIM family, and let $\zeta, \tau > 0$.

$$m^* \;:=\; \min\big\{ m \in \mathbb{Z}_{\geq 0} : \text{ every } f \in \mathcal{F} \text{ is } (m, \zeta, \tau)\text{-well-behaved} \big\}$$

SQ Characterization

There is an efficient SQ algorithm that learns any ζ-noisy function from $\mathcal{F}$ to error $\tau + O(\zeta) + \epsilon$ using

$$N \;=\; d^{\,m^*}\, 2^{\mathrm{poly}(K|\mathcal{Y}|/\epsilon)}$$

samples and time $\mathrm{poly}(N)$. Moreover, no efficient SQ algorithm can achieve error $\tau - O(\zeta)$ with resources $d^{o(m^*)}$.

Corollary

poly(d) learnability occurs if and only if $m^* = O(1)$.

$N = d^{m^*} \mathrm{poly}(K|\mathcal{Y}|/\epsilon)^K$ suffices when y depends only on W.

Results for Well-Studied Function Classes

Function Class	Runtime	Error
General K-MIMs	$d^{O(m^*)}\, 2^{\mathrm{poly}(K\|\mathcal{Y}\|/\epsilon)}$	$\tau + \mathrm{OPT} + \epsilon$
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	$\mathrm{poly}(d)\, 2^{\mathrm{poly}(K/\epsilon)}$	$K\, \tilde{O}(\mathrm{OPT}) + \epsilon$

Subsequent Work & Open Problems

Similar techniques have been used for learning Real-Valued MIMs [Diakonikolas-I-Kane-Ren'25] [Damian-Lee-Bruna'25]. Applications to NNs.

Subsequent Work & Open Problems

Similar techniques have been used for learning Real-Valued MIMs [Diakonikolas-I-Kane-Ren'25] [Damian-Lee-Bruna'25]. Applications to NNs.

Open Problem 1

Is there a $\mathrm{poly}(d, K, 1/\epsilon)$-time learning algorithm for learning MLCs with label noise?

Open Problem 2

Is there a $\mathrm{poly}(d)\, f(K, 1/\epsilon)$-time agnostic learner for intersections of halfspaces with error $O(\mathrm{OPT}) + \epsilon$?

Open Problem 3

Is there an efficient learner for non-Gaussian MIMs?

Subsequent Work & Open Problems

Similar techniques have been used for learning Real-Valued MIMs [Diakonikolas-I-Kane-Ren'25] [Damian-Lee-Bruna'25]. Applications to NNs.

Open Problem 1

Is there a $\mathrm{poly}(d, K, 1/\epsilon)$-time learning algorithm for learning MLCs with label noise?

Open Problem 2

Is there a $\mathrm{poly}(d)\, f(K, 1/\epsilon)$-time agnostic learner for intersections of halfspaces with error $O(\mathrm{OPT}) + \epsilon$?

Open Problem 3

Is there an efficient learner for non-Gaussian MIMs?

Thank you!
Questions?