

Embedding Hard Learning Problems in Gaussian Space

Adam Klivans
UT Austin

Pravesh Kothari
UT Austin

Lower Bounds for Learning

A lot of work on proving lower bounds on learnability of natural concept classes....

[Klivans-Sherstov06],[Klivans-Sherstov10],[Diakonikolas et al.11],[Feldman-Kanade12],[Feldman12],[Berthet-Rigollet13],[Klivans et. al.13],[Daniely et. al.4]

Lower Bounds for Learning

Not many hardness results that apply in widely studied settings for designing algorithms...

- **Distribution Specific**: Uniform, Gaussian...
- **Representation Independent**: Hypothesis in any efficiently computable representation

Lower Bounds for Learning

Not many hardness results that apply when

- **Distribution Specific**: Uniform, Gaussian...
- **Representation Independent**: Hypothesis in any efficiently computable representation

Fast learning algorithms under challenging noise models

- make distributional assumptions: uniform, spherical Gaussian...
- allow output hypothesis any efficiently computable representation. e.g. learn halfspaces via PTF hypotheses

Learning on Gaussian Distribution

A strong simplifying assumption in designing learning algorithms...

γ_n : spherical multivariate Gaussian with variance I in each direction

Hardness of learning for natural concept classes remains open in this setting!

This Work

Lower Bounds for Agnostic
Learning on the Gaussian Distribution

Our Results: I

Agnostically Learning Halfspaces

learner gets random examples from an *arbitrary* f .

Must return hypothesis h :

$$\Pr_D[f(x) \neq h(x)] \leq \text{opt} + \epsilon$$

Agnostically Learning Halfspaces

learner gets random examples from D and f . error of the

closest halfspace to f

Must return hypothesis h

$$\Pr_D[f(x) \neq h(x)] \leq \text{opt} + \epsilon$$

Agnostically Learning Halfspaces

learner gets random examples from *arbitrary* f .

accuracy
parameter

Must return hypothesis h .

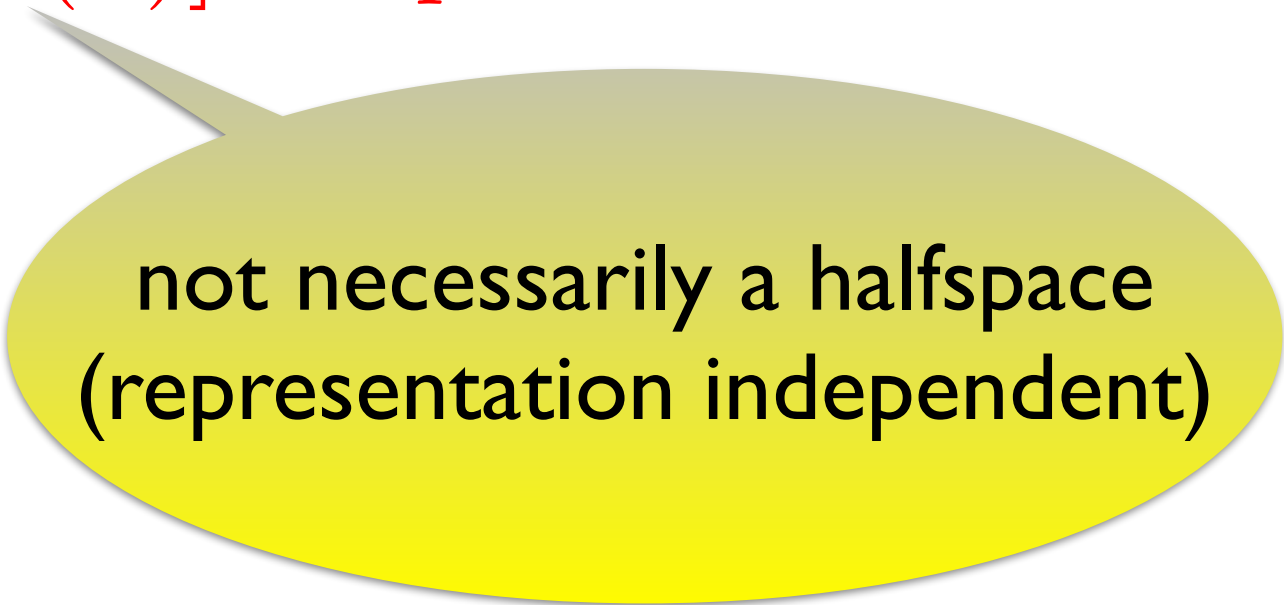
$$\Pr_D[f(x) \neq h(x)] \leq \text{opt} + \epsilon$$

Agnostically Learning Halfspaces

learner gets random examples from an *arbitrary* f .

Must return hypothesis h :

$$\Pr_D[f(x) \neq h(x)] \leq \text{opt} + \epsilon$$



not necessarily a halfspace
(representation independent)

Agnostically Learning Halfspaces

learner gets random examples from an *arbitrary* f .

Must return hypothesis h :

$$\Pr_D[f(x) \neq h(x)] \leq \text{opt} + \epsilon$$



distribution over
example points

Agnostically Learning Halfspaces

[Kalai et al. 2005]

$n^{O(1/\epsilon^2)}$ time algorithm on the uniform distribution
on the **Boolean hypercube** $\{-1, 1\}^n$

[Kalai et al. 2005], [Diakonikolas et al. 2009]

$n^{O(1/\epsilon^2)}$ time algorithm on the uniform distribution
on spherical **Gaussian distribution** γ_n .

Agnostically Learning Halfspaces

[Kalai et al. 2005]

$n^{O(1/\epsilon^2)}$ time algorithm on the uniform distribution
on the **Boolean hypercube**

optimal assuming hardness of learning
parities with noise [Kalai et al.05]

[Kalai et al. 2005], [Diakonikolas et al. 2009]

$n^{O(1/\epsilon^2)}$ time algorithm on the uniform distribution
on spherical **Gaussian distribution**.

Agnostically Learning Halfspaces

[Kalai et al. 2005]

$n^{O(1/\epsilon^2)}$ time algorithm on the uniform distribution
on the **Boolean hypercube**

optimal assuming hardness of learning
parities with noise [Kalai et al.05]

[Kalai et al. 2005], [Diakonikolas et al. 2009]

$n^{O(1/\epsilon^2)}$ time algorithm on the uniform distribution
on spherical **Gaussian distribution**.

?

Agnostically Learning Halfspaces

Is there a $\text{poly}(n, 1/\epsilon)$ time algorithm for agnostically learning halfspaces on γ_n ?

Agnostically Learning Halfspaces

Is there a $\text{poly}(n, 1/\epsilon)$ time algorithm for agnostically learning halfspaces on γ_n ?

NO!

Lower bound of $n^{\Omega(\log(1/\epsilon))}$

This Work

Our Results: 2

Agnostically Learning Sparse* Polynomials

[Andoni et al. 2013]

PAC learner for s -sparse, degree- d
real valued polynomials on the Gaussian
distribution in time $C_d \cdot \text{poly}(s, n, 1/\epsilon)$.
also succeeds under random Gaussian noise

*sparsity = number of monomials

Agnostically Learning Sparse* Polynomials

[Andoni et al. 2013]

PAC learner for s -sparse, degree- d
real valued polynomials on the Gaussian
distribution in time $C_d \cdot \text{poly}(s, n, 1/\epsilon)$.
also succeeds under random Gaussian noise

On the uniform distribution: a generalization of the junta problem!

*sparsity = number of monomials

Agnostically Learning Sparse* Polynomials

no noise ; opt = 0

[Andoni et al. 2013]

PAC learner for *s*-sparse, degree-*d*
real valued polynomials on the Gaussian
distribution in time $C_d \cdot \text{poly}(s, n, 1/\epsilon)$.
also succeeds under random Gaussian noise

*sparsity = number of monomials

Agnostically Learning Sparse* Polynomials

Is there a $C_d \cdot \text{poly}(s, n, 1/\epsilon)$ time algorithm
for *agnostically* learning s -sparse degree d
polynomials on γ_n ?

Agnostically Learning Sparse* Polynomials

Is there a $C_d \cdot \text{poly}(s, n, 1/\epsilon)$ time algorithm
for *agnostically* learning s -sparse degree d
polynomials on γn ?

NO!

This Work

Hardness Assumption

(a hard learning problem on the
uniform distribution on the Boolean hypercube)

Learning Parities With Noise



gets examples $\{(x_i, y_i)\}_{i=1}^m$

Learning Parities With Noise



gets examples $\{(x_i, y_i)\}_{i=1}^m$

$$x_i \sim_U \{-1, 1\}^n$$

Learning Parities With Noise



gets examples $\{(x_i, y_i)\}_{i=1}^m$

$$x_i \sim_U \{-1, 1\}^n$$

$$y_i = \begin{cases} \chi_S(x_i) & \text{w.p. } 1-\eta \\ -\chi_S(x_i) & \text{w.p. } \eta \end{cases}$$

Learning Parities With Noise



gets examples $\{(x_i, y_i)\}^m$

$$x_i \sim_U \{-1, 1\}$$

$$y_i = \begin{cases} \chi_S(x_i) & \text{w.p. } 1-\eta \\ -\chi_S(x_i) & \text{w.p. } \eta \end{cases}$$

$\eta =$
noise rate

Learning Parities With Noise



gets examples $\{(x_i, y_i)\}_{i=1}^m$

$$x_i \sim_U \{-1, 1\}^n$$

$$y_i = \begin{cases} \chi_S(x_i) & \text{w.p. } 1-\eta \\ -\chi_S(x_i) & \text{w.p. } \eta \end{cases}$$

χ_S : parity function on coordinates in

$$S \subseteq [n]$$

Learning Parities With Noise



gets examples $\{(x_i, y_i)\}_{i=1}^m$

$$x_i \sim_U \{-1, 1\}^n$$
$$y_i = \begin{cases} \chi_S(x_i) & \text{w.p. } 1-\eta \\ -\chi_S(x_i) & \text{w.p. } \eta \end{cases}$$

χ_S : parity function on coordinates in $S \subseteq [n]$



learner must return the set S

Learning ^{sparse} Parities With Noise



gets examples $\{(x_i, y_i)\}_{i=1}^m$

$$x_i \sim_U \{-1, 1\}^n$$
$$y_i = \begin{cases} \chi_S(x_i) & \text{w.p. } 1-\eta \\ -\chi_S(x_i) & \text{w.p. } \eta \end{cases}$$

χ_S : parity function on coordinates in $S \subseteq [n]$

$$|S| \leq k$$



learner must return the set S

Learning Sparse Parities With Noise

Brute Force Search: $n^k \cdot \text{poly}(1/\eta)$

Learning Sparse Parities With Noise

Brute Force Search: $n^k \cdot \text{poly}(1/(1 - 2\eta))$

Current Best: $n^{0.8k} \cdot \text{poly}(1/(1 - 2\eta))$

[Valiant 2012]

Learning Sparse Parities With Noise

Brute Force Search: $n^k \cdot \text{poly}(1/\eta)$

Current Best: $n^{0.8k}$ [Valiant 2012]

HARD!

Hardness Assumption

There is no $n^{o(k)} \cdot \text{poly}(1/\eta)$ algorithm to learn sparse parities with noise.

Previously used in [Kalai et al.05],[Feldman et al.13]

captures the hardness of the well known junta problem in learning theory

Hardness of agnostic learning of
halfspaces on the Gaussian
distribution.

Idea

Use agnostic learner for halfspaces on the Gaussian distribution

to

learn sparse parities with noise on the uniform distribution on the Boolean hypercube.

same as the high level idea of [Kalai et al. 2005]

Proof Idea of [Kalai et al.05]

Use agnostic learner for halfspaces on the **uniform distribution** to learn sparse parities with noise on the uniform distribution on the **Boolean hypercube**.

Proof Idea of [Kalai et al.05]

$$\{(x^i, y^i)\}_{i=1}^m$$

Examples labeled by noisy parity.

Observation:

For $k \in [n]$ modify the examples by dropping coordinate k from each point x^i and keeping the same label.

Reduction from [Kalai et al.05]

$$\{(x^i, y^i)\}_{i=1}^m$$

Examples labeled by noisy parity.

Observation: For $k \in [n]$ modify the examples by dropping coordinate k from each point x^i and keeping the same label.

If $k \in S$ then, the resulting labels are uniformly random and independent of the example points.

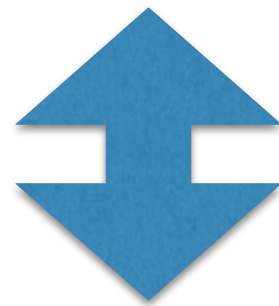
If $k \notin S$ then, the resulting labeled examples are labeled by the same noisy parity seen as a function on $n-1$ bits.

Reduction from [Kalai et al.05]

$$\{(x^i, y^i)\}_{i=1}^m$$

Examples labeled by noisy parity.

Use agnostic learner for halfspaces as a distinguisher to find out if a given variable is relevant or not.



Use that Majority and Parity functions are noticeably correlated

Our Proof

Extend the idea from [Kalai et al.05]

1. Map examples from the Boolean hypercube into Gaussian space.
2. Obtain test for a relevant variable using the agnostic learner for halfspaces.

Our Proof

Extend the idea from [Kalai et al.05]

1. Map examples from the Boolean hypercube into Gaussian space.

A Simple Fact

Half Normal Distribution

“distributed as the absolute value of a standard Gaussian”

A Simple Fact

Half Normal Distribution

“distributed as the absolute value of a standard Gaussian”

h : vector of independent half normals $\in \mathbb{R}_+^n$

x : uniformly random point from $\{-1, 1\}^n$

A Simple Fact

Half Normal Distribution

“distributed as the absolute value of a standard Gaussian”

h : vector of independent half normals $\in \mathbb{R}_+^n$

x : uniformly random point from $\{-1, 1\}^n$

Coordinate-wise Product: $h \circ x$

A Simple Fact

Half Normal Distribution

“distributed as the absolute value of a standard Gaussian”

h : vector of independent half normals $\in \mathbb{R}_+^n$

x : uniformly random point from $\{-1, 1\}^n$

Coordinate-wise Product: $h \circ x$

$$h \circ x \sim \mathcal{N}(0, 1)^n$$

Mapping Labeled Examples

Labeled Examples from
the Uniform Distribution

$$\{(x^i, y^i)\}_{i=1}^m$$

vectors of independent
half-normals

$$h^1, h^2, \dots, h^m$$

Mapping Labeled Examples

Labeled Examples from
the Uniform Distribution

$$\{(x^i, y^i)\}_{i=1}^m$$

vectors of independent
half-normals

$$h^1, h^2, \dots, h^m$$

labeled examples on
the Gaussian Distribution

$$\{(h^i \circ x^i, y^i)\}_{i=1}^m$$

Mapping Labeled Examples

Labeled Examples from
the Uniform Distribution

$$\{(x^i, y^i)\}_{i=1}^m$$

vectors of independent
half-normals

$$h^1, h^2, \dots, h^m$$

labeled examples
the Gaussian

Gaussian Lift

$$\}_{i=1}^m$$

2. Obtain test for a relevant variable using the agnostic learner for halfspaces.

Correlation Lower Bound with Lifted Noisy Parity

1. Doesn't follow from the correlation with parity and majority on the uniform distribution.
2. Exponentially small in the size of the unknown parity unlike the inverse polynomial bound on the uniform distribution on the hypercube.

This is the technical part of the proof.

Correlation Lower Bound with Lifted Noisy Parity

Lemma

$$\langle \text{Maj}_S, \chi_S \rangle_{\gamma^n} = 2^{-\Theta(|S|)}$$

Correlation Lower Bound with Lifted Noisy Parity

Lemma

$$\langle \text{Maj}_S, \chi_S \rangle_{\gamma^n} = 2^{-\Theta(|S|)}$$

Compute the quantity as a limit of an easier expression.

Hardness of Agnostic Learning Halfspaces

Any agnostic learner for halfspaces on the Gaussian distribution requires $n^{\Omega(\log(1/\epsilon))}$ time.

Summary

1. There is no poly-time agnostic learner for halfspaces on the Gaussian distribution - “strongest” distributional assumption.
2. There is poly-time agnostic learner for sparse polynomials on the Gaussian distribution.

Open Question

Best agnostic algorithm for
learning halfspaces

$$n^{O(1/\epsilon^2)}$$

Best hardness

$$n^{\Omega(\log(1/\epsilon))}$$

Open Question

Best agnostic algorithm for
learning halfspaces

$$n^{O(1/\epsilon^2)}$$

Best hardness

$$n^{\Omega(\log(1/\epsilon))}$$

Close the gap!