# SQ Lower Bounds for Learning Halfspaces with Massart Noise

Ilias Diakonikolas, **Daniel Kane**

# Learning Halfspaces

**Definition:** A linear threshold function (LTF) is a function $f: \mathbb{R}^d \to \{-1,1\}$ given by

$$f(x) = \mathrm{sgn}(v \cdot x - t)$$

for some $v \in \mathbb{R}^d$, $t \in \mathbb{R}$.

**Problem:** For some unknown distribution D on $\mathbb{R}^d$ and unknown LTF f, given samples (x,y) with $x \sim D$, $y = f(x)$, learn a hypothesis h so that:

$$\Pr_{x \sim D}(h(x) \neq f(x)) < \varepsilon.$$

**Theorem [Maass-Turan'94]:** This problem can be solved in poly(d/ε) samples and time.

# Noise

It is unrealistic to assume that our data is 100% accurate.

- Assume some (small) probability that $y \neq f(x)$.
- What kinds of learning are still possible?

<u>First question:</u> What noise model to consider?

# Agnostic Noise

[Haussler'92, Kearns-Shapire-Sellie'94]

- Allow for arbitrary (uncommon) errors.
- Can no longer hope to perfectly recover f.

**Problem:** For some distribution D on $\mathbb{R}^d$ x {-1,1} and LTF f, let

$$\mathrm{OPT} = \Pr_{(x,y)\sim D}(f(x) \neq y).$$

Given samples (x,y) ~ D, learn a hypothesis h s.t.:

$$\Pr_{x\sim D}(h(x) \neq y) < \mathrm{OPT} + \varepsilon.$$

- This is information-theoretically possible.
- Settle for O(OPT)+ε or poly(OPT)+ε.

# Hardness

**Theorem [Daniely'16]:** Assuming plausible hardness assumptions about random k-XOR, there is no polynomial time algorithm that distinguishes between $OPT = \exp(-\log^{0.99}(d))$ and $OPT = ½ - d^{-0.01}$.

- Cannot get error much better than ½ even if OPT is almost polynomially small.
- Result also implies SQ lower bounds.
- Agnostic noise too hard.
- Want an easier noise model.

# Random Noise

[Angluin-Laird'88]

**Definition:** A sample with random classification noise (RCN) at rate $\eta$ gives a sample (x,y) with x ~ D and y is:

| | |
|---|---|
| f(x) | with probability 1-$\eta$ |
| - f(x) | with probability $\eta$ |

**Theorem [Blum-Frieze-Kannan-Vempala'96]:** We can learn an LTF with RCN to error $\eta$+ε ( = OPT + ε) in poly(d/ε) samples and time.

# Proof Idea

RCN behaves very well with SQ algorithms.

- $\mathbb{E}_{RCN}[G(x,y)] = (1-\eta)\mathbb{E}[G(x,y)] + \eta\mathbb{E}[G(x,-y)]$

Given h, find function G so that for all x:

$$(1-\eta)G(x,-1) + \eta G(x,1) = h(x,-1)$$

$$(1-\eta)G(x,1) + \eta G(x,-1) = h(x,1)$$

Then

$$\mathbb{E}_{RCN}[G(x,y)] = \mathbb{E}[h(x,y)].$$

So you can simulate noiseless queries.

# Better Noise Models

- RCN is too predictable.
  - Can exactly cancel noise in expectations.
  - Leads to unrealistic algorithms.
- For real problems, we would expect that some examples are more likely to be misclassified than others.
  - This would mess with our algorithms.

# Massart Noise

**Definition:** A sample with Massart noise at rate $\eta$ < 1/2 gives a sample (x,y) with x ~ D and for some function $\eta$(x) < $\eta$, y is:

| | |
|---|---|
| f(x) | with probability 1-$\eta$(x) |
| - f(x) | with probability $\eta$(x) |

**Theorem [Diakonikolas-Gouleakis-Tzamos'19]:** We can learn an LTF with Massart noise to error $\eta$+ε in poly(d/ε) samples and time.

# Error Rates

- For RCN, OPT = $\eta$
  - error $\eta$+ε is best possible.
- For Massart noise OPT might be much smaller.
- Can we learn to error OPT+ε?

**Theorem [Chen-Koehler-Moitra-Yau'20]:** There is no SQ algorithm with polynomial accuracy/queries that learn an LTF with Massart noise to error OPT+o(1) for all OPT and $\eta$.

# What Can We Achieve?

Can we get O(OPT)? Poly(OPT)?
What if we assume that OPT or $\eta$ is small?

**<u>Question:</u>** When learning halfspaces with Massart noise, what is the best error that can be learned efficiently as a function of OPT and $\eta$?

# Hardness

**Theorem [Diakonikolas-K]:** There is no polynomial query/accuracy statistical query algorithm that learns an LTF with Massart noise rate $\boldsymbol{\eta = 1/3}$ to error better than **1/polylog(d)** even when guaranteed that $\mathbf{OPT < \exp(-\log^{0.99}(d))}$.

- Size of OPT comparable to Daniely.
- Achievable error worse (would like $\eta+\varepsilon$).

# SQ Lower Bounds

Recall the basic result for proving SQ lower bounds:

**Proposition:** Let A be a distribution on $\mathbb{R}$ that matches k moments with N(0,1) to error $\nu$
Any SQ algorithm that distinguishes N(0,I) from $P^A_v$ either:

- Makes queries of error at most
  $$\tau = d^{-ck}\,\chi^2(A) + \nu^2$$
- Makes at least $\exp(d^c)\,\tau / \chi^2(A)$ queries

Needs to be able to deal inexact moment matching.

# Lower Bounds for Functions

The old techniques are great for showing that it is hard to learn distributions x. But our algorithm sees (x,y) pairs and y is not remotely Gaussian.

Instead we make $(x|y=1)$ and $(x|y=-1)$ hard to distinguish.

It turns out this is enough.

# New Lower Bound

**Proposition:** Let A, B be distributions on $\mathbb{R}$ that matches k moments with N(0,1) to error $\nu$. Let $p \in (0,1)$. For unit vector v, let $P_v$ be the distribution on $\mathbb{R}^d \times \{-1,1\}$ that returns $(P^A_v, 1)$ with probability p and $(P^B_v, -1)$ with probability 1-p. Then any SQ algorithm that distinguishes $N(0,I) \times \{-1,1\}$ from $P_v$ either:

- Makes queries of error at most

$$\tau = d^{-ck} (\chi^2(A)+\chi^2(B)) + \nu^2$$

- Makes at least $\exp(d^c)\, \tau / (\chi^2(A)+\chi^2(B))$ queries

Need $P_v$ to be an LTF with Massart noise.

# Problem

Any distribution A (approximately) matching $O(1/\varepsilon^2)$ moments with a Gaussian has

$$\Pr(A>t) = \Pr(G>t)+O(\varepsilon).$$

We cannot afford for this to happen for both $(x|y=1)$ and $(x|y=-1)$.

To solve this, we will need to fool LTFs in some more complicated space.

# Polynomial Threshold Functions

**<u>Definition:</u>** A degree-k <u>Polynomial Threshold Function</u> (PTF) is a function of the form

$$f(x) = \mathrm{sgn}(p(x))$$

for p some polynomial of degree at most k.

**<u>Note:</u>** a PTF is an LTF in the monomials of x.

- $V_k(x)$ = (degree-k monomials of x) $\in \mathbb{R}^N$.
- $f(x) = g(V_k(x))$ for some LTF g.

**<u>Need to Show:</u>** cannot learn a degree-k PTF in poly(N) = poly($d^k$) samples/accuracy.

# Need

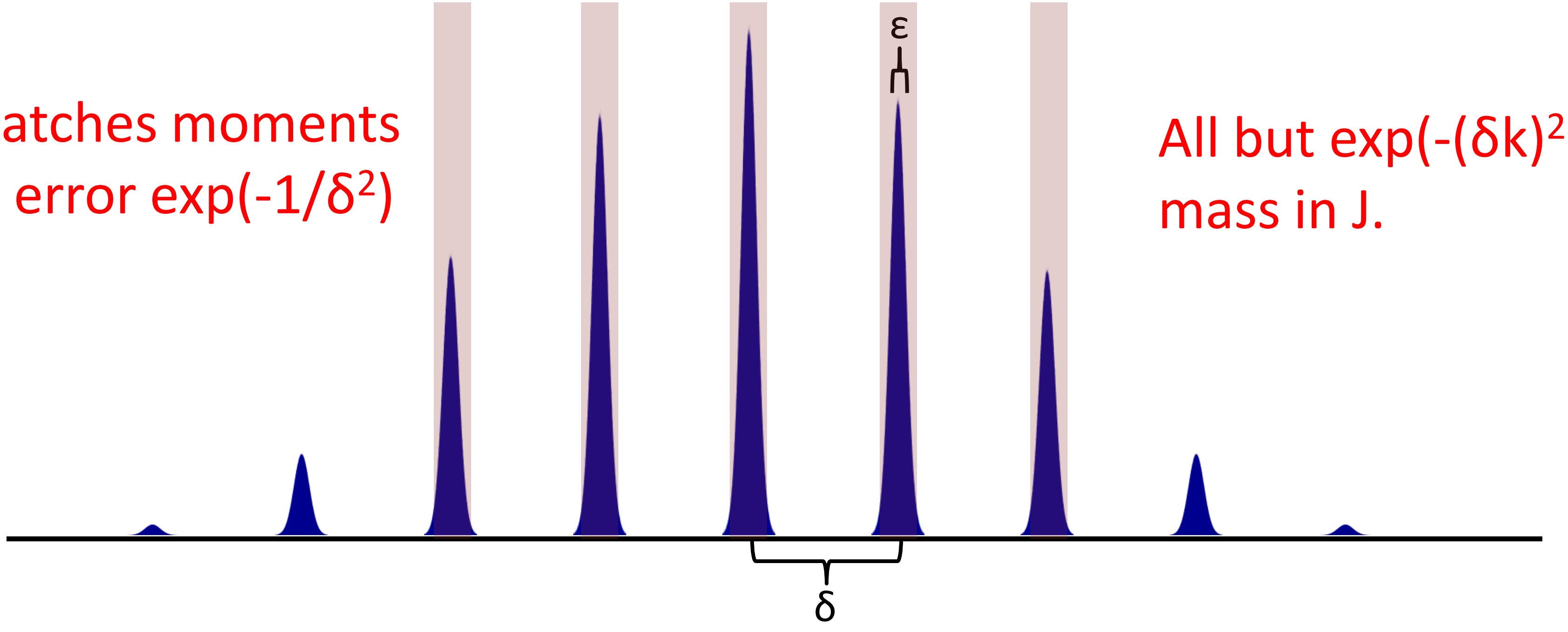
This construction needs distributions A and B and J a union of k/2 intervals so that:

- A and B approximately match $\omega(k)$ moments with $N(0,1)$.
- All but OPT of the mass of B is supported on J
- All but OPT of the mass of A is supported on $J^c$
- $B > 2A$ on J
- $A > 2B$ on $J^c$

# B Construction

- B is a net of Gaussians.
- J is k/2 intervals around peaks.

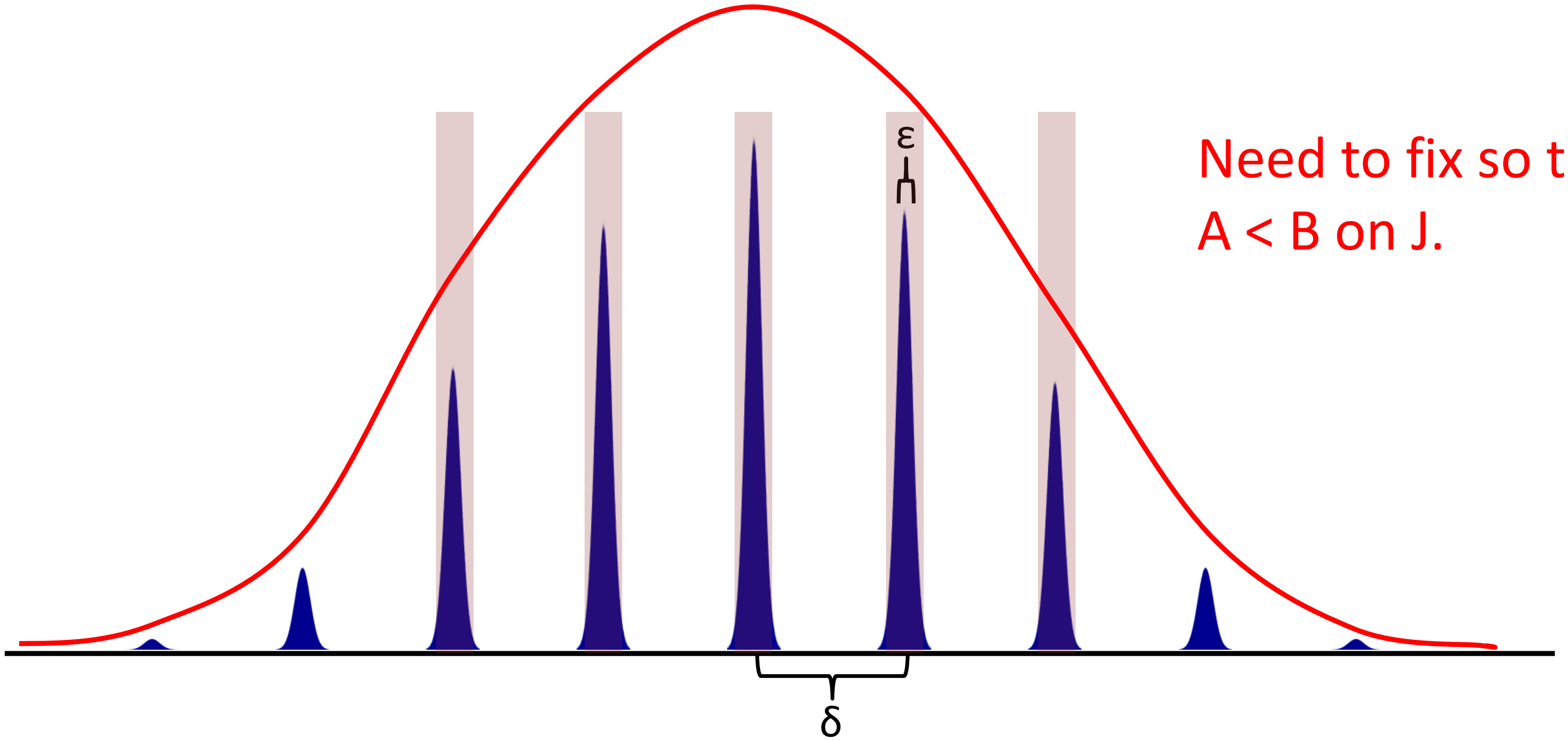
Matches moments to error $\exp(-1/\delta^2)$




All but $\exp(-(\delta k)^2)$ mass in J.

# A Construction

- Start with a taller Gaussian.
  - Matches moments exactly
  - Bigger than B on $J^c$.




Need to fix so that A < B on J.

# Fix

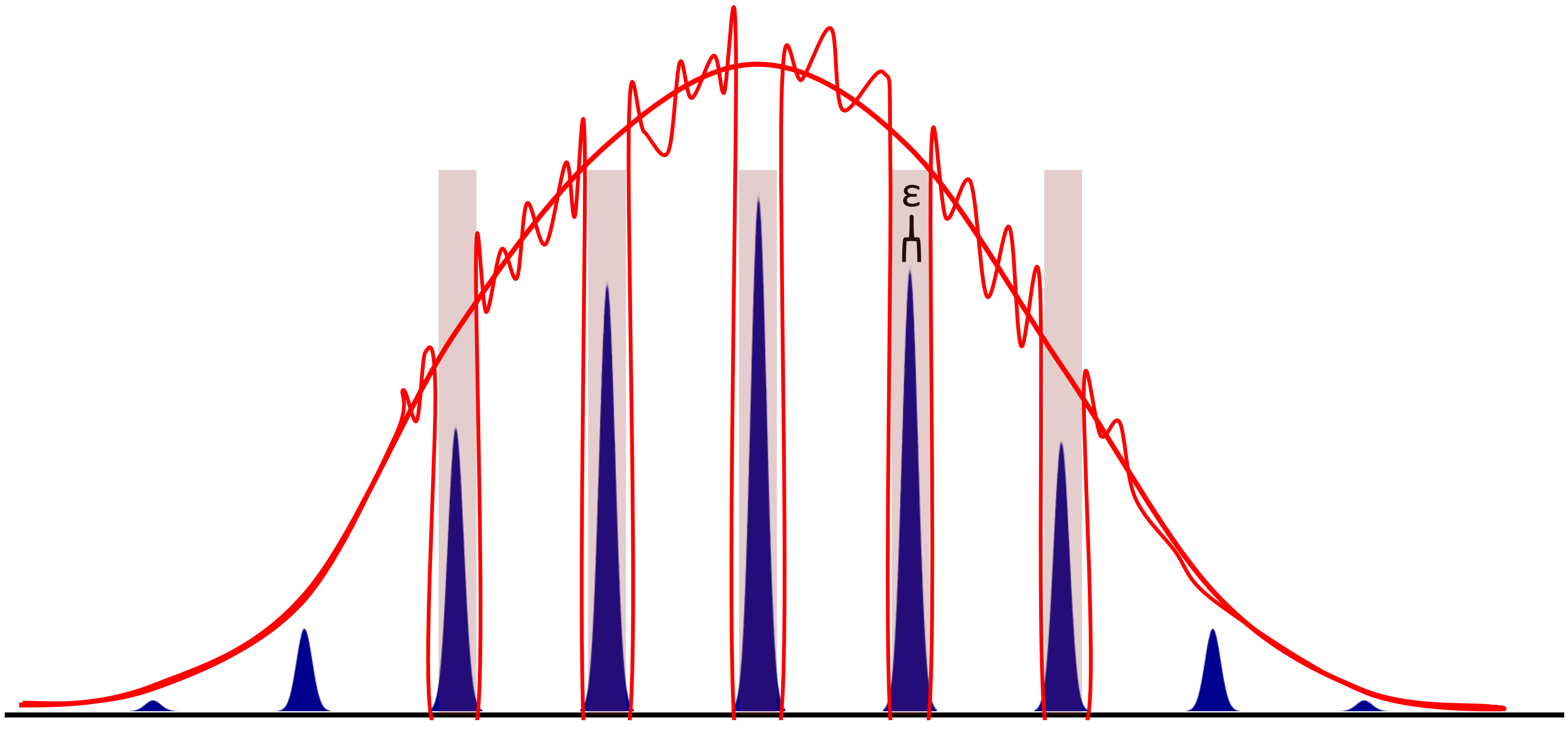
Move mass of A on J off of it without changing the first m moments.

**Lemma:** Let D be a distribution on [-1,1] that is approximately uniform. Let $c \ll 1/m^2$. There exists a distribution D' approximately uniform on [-1,1] \ [-c,c] so that D' and D match m moments.

**Proof idea:** modify the pdf by a polynomial.

# Fix

- Apply modification about each interval in J.



# Parameters

- $N = d^k$, so $k \approx \log(N)$
- Need $\exp(1/\delta^2) >> \text{complexity} >> N$
  - $\delta << k^{-1/2}$
  - $\text{OPT} \approx \exp(-(k\delta)^2) \approx \exp(-k) \approx$ almost $1/\text{poly}(N)$
- $\varepsilon \approx$ Interval width $\approx 1/m^2 << 1/k^2$
  - Need A to be $\delta/\varepsilon$ more mass than B.
  - $p \approx \varepsilon/\delta$
  - Can learn to error $\varepsilon/\delta \approx 1/\text{polylog}(N)$

# Improvement

A more recent refinement of this technique shows that it is hard to get better than constant error even for very small OPT.

# Conclusions

- Can learn to error $\eta$+ε with Massart noise.
- Cannot do much better even if OPT is quite small.
- SQ Lower bounds are a useful tool for getting evidence of hardness for function learning problems.

# Further Work

- Get similar results via reduction from some standard hard problem.
- Get lower bounds for learning other linear models like ReLUs.