

Agnostic learning under permutation invariant distributions

1	2	3	4	5
6	7	8		
9	10			

Karl Wimmer

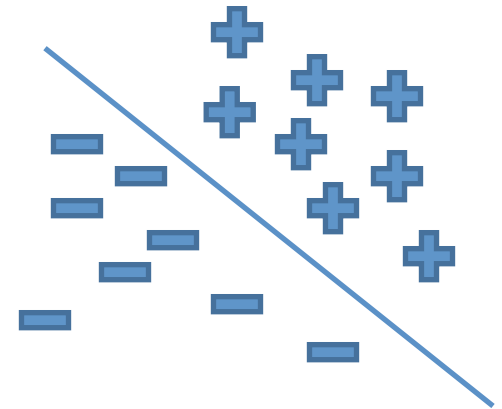
Simons Institute & Duquesne University

Background

Here, we consider ***agnostic learning***, with a special focus on ***linear threshold functions***, over ***finite discrete domains***.

We say that $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ is a ***linear threshold function (LTF)*** if there exists weights $\{w_i\}_{i=0}^n$ such that

$$f(x) = \text{sgn}(w_0 + \sum_{i=1}^n w_i x_i)$$



Background

Here, we consider ***agnostic learning***, with a special focus on ***linear threshold functions***, over ***finite discrete domains***.

We say that C is ***agnostically learnable with respect to*** D if there is an efficient algorithm that outputs h such that

$$\Pr_{\mathbf{x} \sim D}[t(\mathbf{x}) \neq h(\mathbf{x})] \leq \min_{f \in C} \Pr_{\mathbf{x} \sim D}[t(\mathbf{x}) \neq f(\mathbf{x})] + \epsilon$$

for a target function t .

Background

[KlivansO'DonnellServedio02] showed that intersections of k LTF's are learnable in time $n^{O(k^2/\epsilon^2)}$. (Uniform on $\{-1, 1\}^n$.)

The algorithm [LinialMansourNisan91] is simple regression: solve

$$\begin{aligned} & \min \|t - p\|_2^2 \\ & \text{s.t. } \deg(p) \leq d \end{aligned}$$

with $d = O(k^2/\epsilon^2)$. Output $h = \text{sgn}(p)$

Background

[KKMS05] showed that intersections of k LTF's are ***agnostically*** learnable in time $n^{O(k^4/\epsilon^4)}$. (Uniform on $\{-1, 1\}^n$.)

The algorithm [KKMS05] is simple regression: solve

$$\begin{aligned} & \min \|t - p\|_1 \\ & s.t. \deg(p) \leq d \end{aligned}$$

with $d = O(k^4/\epsilon^4)$. Output $h = \text{sgn}(p)$.

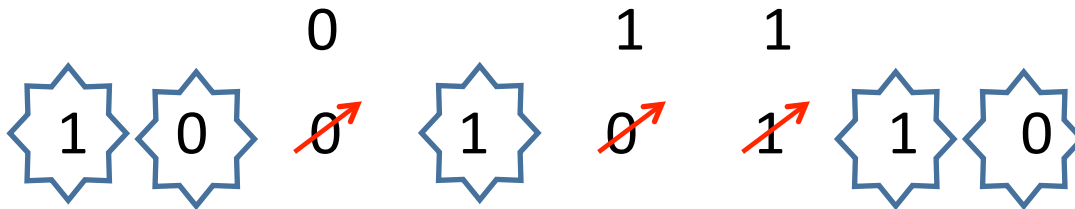
These proofs use the ***noise sensitivity method***.

Background

$$f : \{-1, 1\}^n \rightarrow \{-1, 1\}$$
$$\text{Inf}^{(i)}(f) = \Pr_{\mathbf{x}}[f(\mathbf{x}) \neq f(\mathbf{x}^{(i)})]$$

$$\text{NS}_{\delta}(f) = \Pr_{\mathbf{x}}[f(\mathbf{x}) \neq f(N_{\delta}(\mathbf{x}))]$$

$N_{\delta}(\mathbf{x})$



Independently mark each coordinate **frozen** with probability $1 - \delta$.

Rerandomize the **unfrozen** coordinates

Background

$$\text{NS}_\delta(f) = \Pr_{\mathbf{x}}[f(\mathbf{x}) \neq f(N_\delta(\mathbf{x}))]$$

Key fact: $\text{NS}_\delta(f) = \sum_{S \subseteq [n]} \left(\frac{1}{2} - \frac{1}{2}(1 - 2\delta)^{|S|}\right) \hat{f}(S)^2$

where $f = \sum_{S \subseteq [n]} \hat{f}(S) x_S$

Parseval: $\sum_S \hat{f}(S)^2 = 1$ for Boolean functions

The distance from $\frac{1}{2}$ for the multiplier at degree $d = |S|$ decays at an **exponential** rate in the degree.

Low noise sensitivity \rightarrow lots of mass on low-degree terms \rightarrow
Low degree approximator \rightarrow agnostic learning via regression

[KKMS05]

Background

$$\text{NS}_\delta(f) = \Pr_{\mathbf{x}}[f(\mathbf{x}) \neq f(N_\delta(\mathbf{x}))]$$

$$f : \{-1, 1\}^n \rightarrow \{-1, 1\}$$

[Peres]: If f is an LTF, then $\text{NS}_\delta(f) \leq \sqrt{\delta}$.

Union bound: If f depends on k LTF's, then $\text{NS}_\delta(f) \leq k\sqrt{\delta}$.

Setting $\delta = \epsilon^4/k^4$ gives $\text{NS}_\delta(f) \leq O(\epsilon^2)$, which implies a good approximator in ℓ_1 distance (via ℓ_2 distance).

Works in product distributions too. [BlaisO'DonnellW.08]

Background

$$\text{NS}_\delta(f) = \Pr_{\mathbf{x}}[f(\mathbf{x}) \neq f(N_\delta(\mathbf{x}))]$$

$$f : \{-1, 1\}^n \rightarrow \{-1, 1\}$$

But all this is really only known for (essentially) product distributions.

Setting $\delta = \epsilon/\sqrt{n}$ gives $\text{NS}_\delta(f) \leq O(\epsilon^2)$, which implies a good approximator in ℓ_1 distance (via ℓ_2 distance).

Works in product distributions too. [BlaisO'DonnellW.08]

The setup

A ***permutation invariant distribution*** is a distribution D over $\{-1, 1\}^n$ such that $D(x) = D(\sigma x)$ for any permutation σ .

A distribution is permutation invariant if and only if it is a mixture of distributions that are uniform over $\binom{[n]}{k}$.

For learning, we can focus on learning algorithms for the uniform distribution over $\binom{[n]}{k}$.

We unify these cases by appealing to the symmetric group.

The setup

The uniform distribution over $\binom{[n]}{k}$ is interesting, and can be helpful over problems over $\{-1, 1\}^n$.

Example: KKL Theorem over this distribution [O'DonnellW.09]

This leads to optimally weak-learning monotone functions, settling a question of [BlumBurchLangford95].

Related Work [W.]: Friedgut's junta theorem in this domain. (Like this talk, uses representation theory of symmetric group.)

The setup

$$g : \binom{[n]}{k} \rightarrow \mathbb{R}$$

$$g(\quad -1 \quad -1 \quad 1 \quad 1 \quad \dots \quad 1 \quad -1 \quad)$$



$$f(\quad k+32 \quad k+6 \quad 3 \quad k \quad \dots \quad 1 \quad k-1 \quad)$$

This is a $k!(n-k)!$ -to-one mapping,
so the uniform distribution is induced.

$$f : \text{Sym}_n \rightarrow \mathbb{R}$$

Roadmap to agnostic learning

$$f : \text{Sym}_n \rightarrow \mathbb{R}$$

Define $\text{NS}_\delta(f)$, and obtain a nice Fourier representation.

Show that the distance from $\frac{1}{2}$ for the multipliers decays at an **exponential** rate in the degree.

Bound $\text{NS}_\delta(f)$ for LTF's (and functions of LTF's by union bound).

Efficient agnostic learning over permutation invariant distributions follows from [KKMS05].

Representation theory basics

$$f : \text{Sym}_n \rightarrow \mathbb{R}$$

degree

A **representation** is a map $\rho : \text{Sym}_n \rightarrow \mathbb{R}^{d_\rho \times d_\rho}$ such that

$$\rho(\sigma_1 \sigma_2) = \rho(\sigma_1) \rho(\sigma_2)$$

Irreducible: can not be written as $C^{-1} \begin{bmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{bmatrix} C$

Young's Orthogonal Representation (YOR) gives a **complete set of irreducible representations**.

Representation theory basics

$$f : \text{Sym}_n \rightarrow \mathbb{R}$$

degree

A **representation** is a map $\rho : \text{Sym}_n \rightarrow \mathbb{R}^{d_\rho \times d_\rho}$ such that

$$\rho(\sigma_1 \sigma_2) = \rho(\sigma_1) \rho(\sigma_2) \quad \begin{array}{l} \text{contributes } d_\rho^2 \\ \text{many functions} \end{array}$$

Young's Orthogonal Representation (YOR) gives a **complete set of irreducible representations**.

The functions $\{\rho_{ij}\}$ form an orthogonal basis for the vector space of functions $f : \text{Sym}_n \rightarrow \mathbb{R}$.

Representation theory basics

$$f : \text{Sym}_n \rightarrow \mathbb{R}$$

degree

A **representation** is a map $\rho : \text{Sym}_n \rightarrow \mathbb{R}^{d_\rho \times d_\rho}$ such that Not too bad

$$\rho(\sigma_1 \sigma_2) = \rho(\sigma_1) \rho(\sigma_2)$$

$$\hat{f}_\rho = \mathbb{E}[f(\boldsymbol{\sigma}) \rho(\boldsymbol{\sigma})]$$

$$f = \sum_{\rho} d_\rho \text{tr}(\hat{f}_\rho^T \rho(\boldsymbol{\sigma}))$$

Parseval: $f : \text{Sym}_n \rightarrow \{-1, 1\}$ implies $\sum_{\rho} d_\rho \|\hat{f}_\rho\|^2 = 1$ Frobenius norm;
sum sq's of entries

Representation theory basics

$$f : \text{Sym}_n \rightarrow \mathbb{R}$$

degree

A **representation** is a map $\rho : \text{Sym}_n \rightarrow \mathbb{R}^{d_\rho \times d_\rho}$ such that

$$\rho(\sigma_1 \sigma_2) = \rho(\sigma_1) \rho(\sigma_2)$$

Fantastic result: these irreducible representations can be indexed by partitions of n .

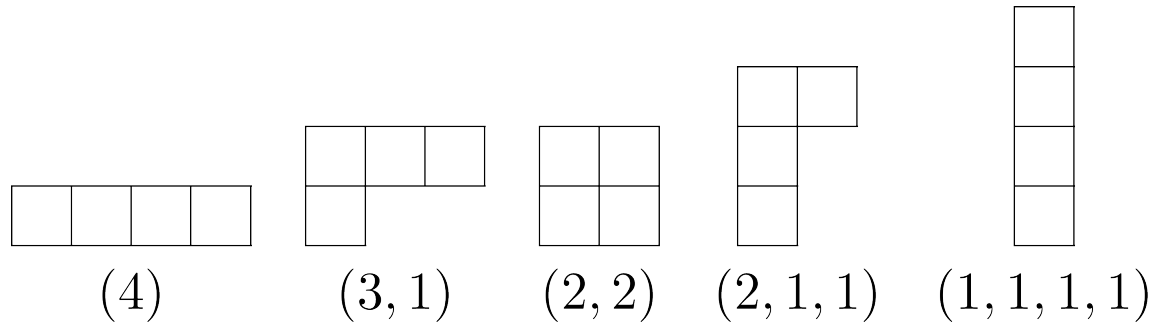
$$\lambda \vdash n, \lambda = (\lambda_1, \lambda_2, \dots, \lambda_r), \lambda_i > 0$$

We will often refer to representations by the corresponding partition.

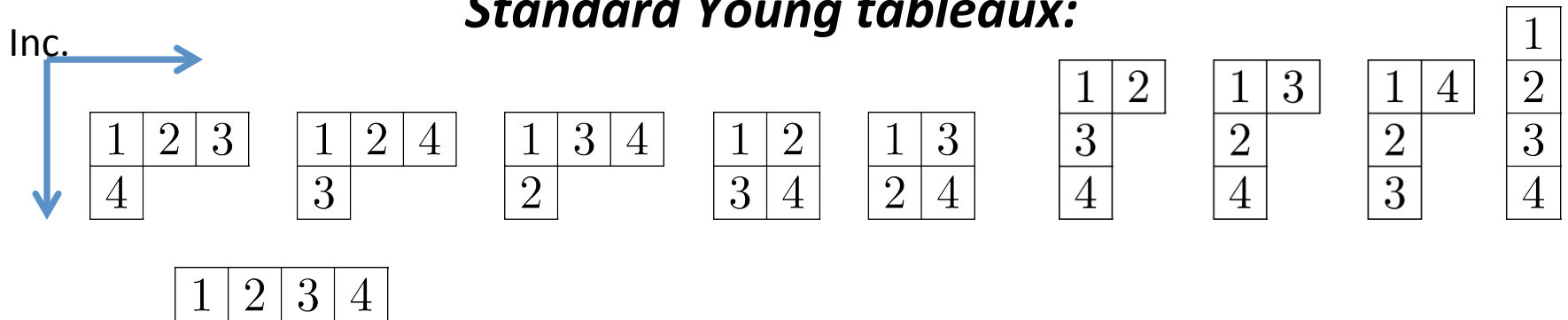
Partitions

$$\lambda \vdash n, \lambda = (\lambda_1, \lambda_2, \dots, \lambda_r), \lambda_i > 0$$

Visualized as *Young diagrams*:



Standard Young tableaux:



Partitions

$$\lambda \vdash n, \lambda = (\lambda_1, \lambda_2, \dots, \lambda_r), \lambda_i > 0$$

Let d_λ be the number of standard Young tableaux with shape λ .

$$d_\lambda = d_{\rho_\lambda}$$

Standard Young tableaux:

Inc.

1	2	3	1	2	4	1	3	4	1	2	1	3	1	4	1	2	3	4
4			3			2			3	4	2	4	3	4	4	4	3	4

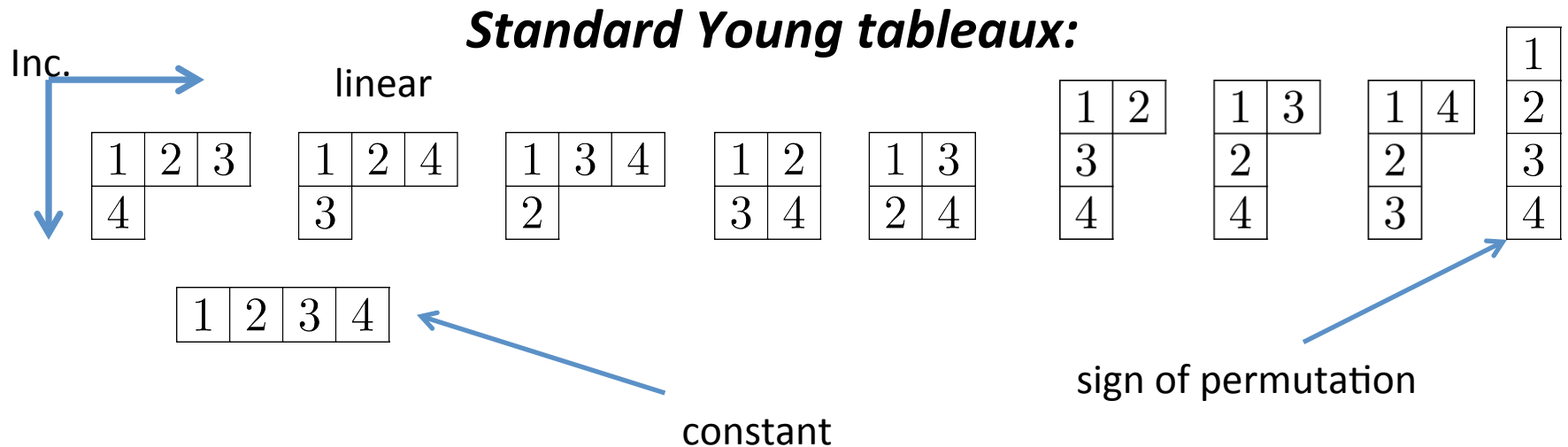
1	2	3	4
---	---	---	---

$$1^2 + 3^2 + 2^2 + 3^2 + 1^2 = 24 = 4!$$

Partitions

$$\lambda \vdash n, \lambda = (\lambda_1, \lambda_2, \dots, \lambda_r), \lambda_i > 0$$

As “polynomials over Sym_n ”, the degree of the representations corresponding to λ have degree $n - \lambda_1$. (Over $\{1[\sigma(i) = j]\}_{1 \leq i, j \leq n}$, or transferring from Sym_n into a subset of $\{0, 1\}^{n^2}$.)



Noise Sensitivity

First step: Influence of a set

$$\text{Inf}^{(S)} = \Pr_{\sigma} [f(\sigma) \neq f(\sigma^{(S)})]$$

$\sigma^{(S)}$ has coordinates in S shuffled

			4		6		8		
			4		8		6		
	3	5	6	1	4	8	7	2	
			6		8	4			
			8		4	6			
			8		6	4			

Noise Sensitivity

k=2
[Diaconis89]

$$\text{Inf}^{(S)} = \Pr_{\sigma} [f(\sigma) \neq f(\sigma^{(S)})]$$

There is a nice expression for the average of these

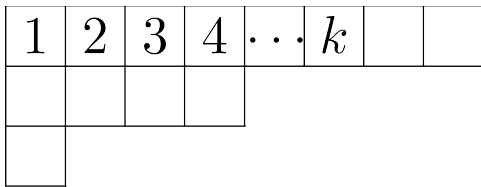
This is nontrivial.

$$\text{avg}_{|S|=k} \text{Inf}^{(S)} = \sum_{\lambda \vdash n} d_{\lambda} \left(\frac{1}{2} - \frac{1}{2} \frac{d_{\lambda/(k)}}{d_{\lambda}} \right) \|\hat{f}_{\lambda}\|^2$$



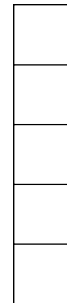
Frobenius norm

$d_{\lambda/(k)}$ is the number of



contributes nothing to influence (constant)

...



contributes lots to influence (sign of permutation)

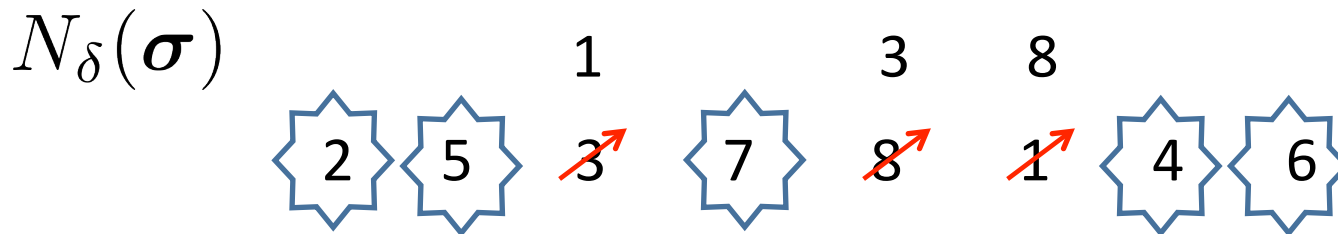
⋮

⋮

Noise Sensitivity

For $f : \text{Sym}_n \rightarrow \{-1, 1\}$, we define noise sensitivity $\text{NS}_\delta(f)$ as:

$$\text{NS}_\delta(f) = \Pr_{\sigma, N_\delta(\sigma)}[f(\sigma) \neq f(N_\delta(\sigma))]$$



Independently mark each coordinate **frozen** with probability $1 - \delta$.

Uniformly shuffle the **unfrozen** coordinates

Trick: pick the number of unfrozen coordinates first.

Noise Sensitivity

For $f : \text{Sym}_n \rightarrow \{-1, 1\}$, we define noise sensitivity $\text{NS}_\delta(f)$ as:

$$\text{NS}_\delta(f) = \Pr_{\sigma, N_\delta(\sigma)}[f(\sigma) \neq f(N_\delta(\sigma))]$$

$N_\delta(\sigma)$



$$\text{NS}_\delta(f) = \mathbb{E}_{\mathbf{k}}[\text{avg}_{|S|=\mathbf{k}} \text{Inf}^S(f)]$$

$$\mathbf{k} \sim \text{Binomial}(n, \delta)$$

Noise Sensitivity

$$\text{NS}_\delta(f) = \mathbb{E}_{\mathbf{k}}[\text{avg}_{|S|=\mathbf{k}} \text{Inf}^S(f)]$$

$$\mathbf{k} \sim \text{Binomial}(n, \delta)$$

$$\text{NS}_\delta(f) = \mathbb{E}_{\mathbf{k}} \left[\sum d_\lambda \left(\frac{1}{2} - \frac{1}{2} \frac{d_{\lambda/(k)}}{d_\lambda} \right) \|\hat{f}_\lambda\|^2 \right]$$

$$\text{NS}_\delta(f) = \sum d_\lambda \left(\mathbb{E}_{\mathbf{k}} \left[\frac{1}{2} - \frac{1}{2} \frac{d_{\lambda/(k)}}{d_\lambda} \right] \right) \|\hat{f}_\lambda\|^2$$

$$\sum_{\lambda \vdash n} d_\lambda \|\hat{f}_\lambda\|^2 = 1$$

Analyze this

Roadmap to agnostic learning

$$f : \text{Sym}_n \rightarrow \mathbb{R}$$



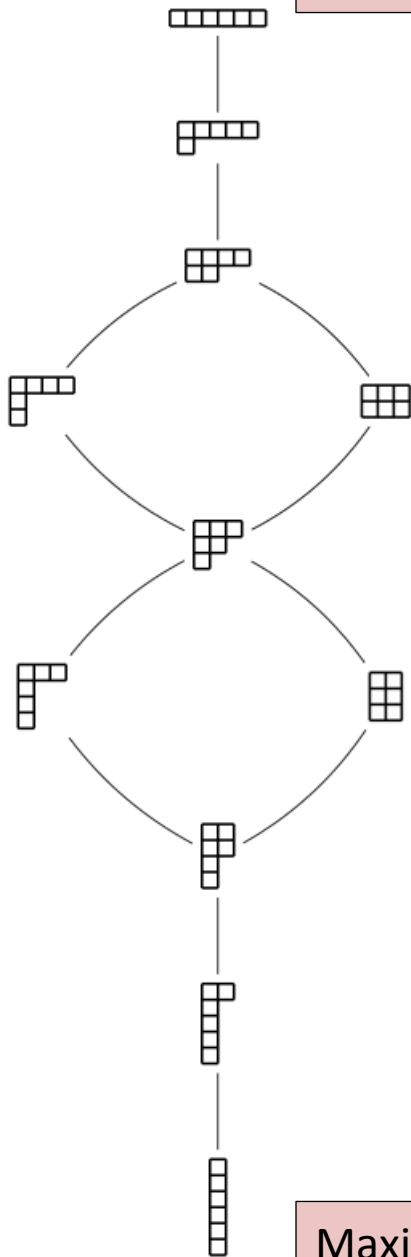
Define $\text{NS}_\delta(f)$, and obtain a nice Fourier representation.

Show that the distance from $\frac{1}{2}$ for the multipliers decays at an **exponential** rate in the degree.

Bound $\text{NS}_\delta(f)$ for LTF's (and functions of LTF's by union bound).

Efficient agnostic learning over permutation invariant distributions follows from [KKMS05].

Minimum noise



Standard Young tableaux

k=2

[Diaconis89]

[Fulton 96] give a handy formula for $\frac{d_{\lambda/(k)}}{d_{\lambda}}$:

$$\sum_{1 \leq i_1 \leq i_2 \leq \dots \leq i_k < \infty} \prod_{j=1}^k (\lambda_{i_j} - k + j)$$

[W.]: $\frac{d_{\lambda/(k)}}{d_{\lambda}} \geq \frac{d_{\beta/(k)}}{d_{\beta}}$ if $\lambda \triangleright \beta$

$$\beta \triangleleft \lambda \iff \forall r \sum_{i=1}^r \lambda_i \geq \sum_{i=1}^r \beta_i$$

Maximum noise

Standard Young tableaux

Theorem [W.]:
$$\frac{d_{\lambda/(k)}}{d_{\lambda}} \geq \frac{d_{\beta/(k)}}{d_{\beta}} \text{ if } \lambda \triangleright \beta$$

$$\lambda \triangleright \beta \leftrightarrow \forall r \sum_{i=1}^r \lambda_i \geq \sum_{i=1}^r \beta_i$$

Proof sketch: By induction, suffices to show the theorem for

$$\lambda = (t + 1, t, t, \dots, t, t - 1) \text{ and } \beta = (t, t, t, \dots, t, t),$$

This case and the induction heavily use the [OO96] formula.

This confirms that “tall” partitions are noisier than “wide” partitions.

Standard Young tableaux

Lemma [W.]:
$$\frac{d_{\lambda/(k)}}{d_{\lambda}} \leq \left(\frac{d_{\lambda/(2)}}{d_{\lambda}} \right)^{k-1}$$

Proof: Induction, manipulatorics, and [OO96].

Standard Young tableaux

Lemma [W.]:
$$\frac{d_{\lambda/(k)}}{d_{\lambda}} \leq \left(\frac{d_{\lambda/(2)}}{d_{\lambda}} \right)^{k-1}$$

Corollary: $\mathbf{k} \sim \text{Binomial}(n, \delta)$

$$\mathbb{E}_{\mathbf{k}} \left[\frac{d_{\lambda/(k)}}{d_{\lambda}} \right] \leq \mathbb{E}_{\mathbf{k}} \left[\left(\frac{d_{\lambda/(2)}}{d_{\lambda}} \right)^{k-1} \right] = \frac{d_{\lambda}}{d_{\lambda/(2)}} \left(1 - \delta \left(1 - \frac{d_{\lambda/(2)}}{d_{\lambda}} \right) \right)^n$$

using the moment generating function for the binomial distribution.

Standard Young tableaux

Lemma [W.]:
$$\frac{d_{\lambda/(k)}}{d_{\lambda}} \leq \left(\frac{d_{\lambda/(2)}}{d_{\lambda}} \right)^{k-1}$$

Corollary: $\mathbf{k} \sim \text{Binomial}(n, \delta)$

$$\mathbb{E}_{\mathbf{k}} \left[\frac{d_{\lambda/(k)}}{d_{\lambda}} \right] \leq \mathbb{E}_{\mathbf{k}} \left[\left(\frac{d_{\lambda/(2)}}{d_{\lambda}} \right)^{k-1} \right] = \frac{d_{\lambda}}{d_{\lambda/(2)}} \left(1 - \delta \left(1 - \frac{d_{\lambda/(2)}}{d_{\lambda}} \right) \right)^n$$

using the moment generating function for the binomial distribution.

Now use the Theorem to maximize this expression.

Noise Sensitivity

$$\text{NS}_\delta(f) = \mathbb{E}_{\mathbf{k}}[\text{avg}_{|S|=\mathbf{k}} \text{Inf}^S(f)]$$

$$\mathbf{k} \sim \text{Binomial}(n, \delta)$$

$$\text{NS}_\delta(f) = \mathbb{E}_{\mathbf{k}} \left[\sum d_\lambda \left(\frac{1}{2} - \frac{1}{2} \frac{d_{\lambda/(\mathbf{k})}}{d_\lambda} \right) \|\hat{f}_\lambda\|^2 \right]$$

$$\text{NS}_\delta(f) = \sum d_\lambda \left(\mathbb{E}_{\mathbf{k}} \left[\frac{1}{2} - \frac{1}{2} \frac{d_{\lambda/(\mathbf{k})}}{d_\lambda} \right] \right) \|\hat{f}_\lambda\|^2$$

$$\sum_{\lambda \vdash n} d_\lambda \|\hat{f}_\lambda\|^2 = 1$$

Standard Young tableaux

Theorem [W.]: $\frac{d_{\lambda/(k)}}{d_{\lambda}} \geq \frac{d_{\beta/(k)}}{d_{\beta}}$ if $\lambda \triangleright \beta$

$$\lambda \triangleright \beta \Leftrightarrow \forall r \sum_{i=1}^r \lambda_i \geq \sum_{i=1}^r \beta_i$$

$$\mathbb{E}_{\mathbf{k}} \left[\frac{d_{\lambda/(k)}}{d_{\lambda}} \right] \leq \mathbb{E}_{\mathbf{k}} \left[\left(\frac{d_{\lambda/(2)}}{d_{\lambda}} \right)^{k-1} \right] = \frac{d_{\lambda}}{d_{\lambda/(2)}} \left(1 - \delta \left(1 - \frac{d_{\lambda/(2)}}{d_{\lambda}} \right) \right)^n$$

Over partitions with $\lambda_1 \leq n - d$, this is maximized for $\lambda = (n - d, d)$.


We get roughly $1 - \frac{d_{\lambda/(2)}}{d_{\lambda}} = d/n$, the expression becomes roughly $\exp(-\delta d)$.

Polynomial degree at least d


Thus the noise exponentially approaches $\frac{1}{2}$ as the degree increases, and low noise sensitivity implies low degree concentration.

Roadmap to agnostic learning

$$f : \text{Sym}_n \rightarrow \mathbb{R}$$



Define $\text{NS}_\delta(f)$, and obtain a nice Fourier representation.



Show that the distance from $\frac{1}{2}$ for the multipliers decays at an **exponential** rate in the degree.

Bound $\text{NS}_\delta(f)$ for LTF's (and functions of LTF's by union bound).

Efficient agnostic learning over permutation invariant distributions follows from [KKMS05].

Linear threshold functions

We say that $f : \text{Sym}_n \rightarrow \{-1, 1\}$ is a **linear threshold function (LTF)** if there exist weights $\{w_{ij}\}_{1 \leq i, j \leq n}$ and θ such that

$$f(\sigma) = \text{sgn} \left(\sum w_{ij} \mathbf{1}[\sigma(i) = j] - \theta \right)$$

 Indicator variables

Equivalently, $f(\sigma) = \text{sgn}(\text{tr}(W^T P(\sigma)) - \theta)$, where W is the matrix of weights, and P represents the permutation matrix encoding.

Theorem [W.]: If $f : \text{Sym}_n \rightarrow \{-1, 1\}$ is an LTF, then

$$\text{NS}_\delta(f) \leq O(\sqrt{\delta})$$

Linear threshold functions

Theorem [W.]: If $f : \text{Sym}_n \rightarrow \{-1, 1\}$ is an LTF, then

$$\text{NS}_\delta(f) \leq O(\sqrt{\delta})$$

Proof: Reduction to hypercube case.

$$\text{NS}_\delta(f) = \Pr_{\sigma, N_\delta(\sigma)}[f(\sigma) \neq f(N_\delta(\sigma))]$$

Assume that $1/\delta = m$ is an integer. Partition the coordinates into m buckets, independently putting each in any bucket with probability $1/m$.

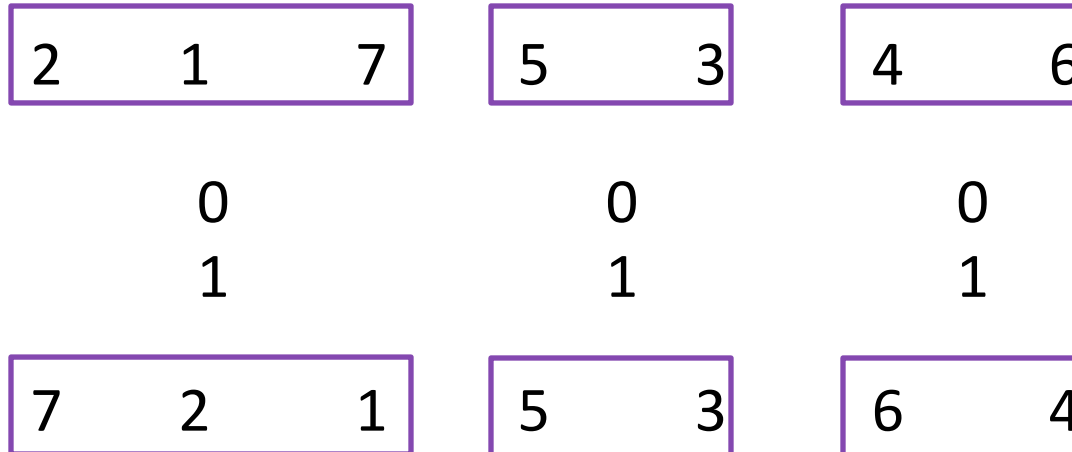
Linear threshold functions

Theorem [W.]: If $f : \text{Sym}_n \rightarrow \{-1, 1\}$ is an LTF, then

$$\text{NS}_\delta(f) \leq O(\sqrt{\delta})$$

Example: $n=7$, $m=3$, and the buckets are $\{1,2,3\}$, $\{4,5\}$, and $\{6,7\}$.

Draw σ .



Draw a permutation that only shuffles coordinates in each bucket.

Linear threshold functions

Theorem [W.]: If $f : \text{Sym}_n \rightarrow \{-1, 1\}$ is an LTF, then

$$\text{NS}_\delta(f) \leq O(\sqrt{\delta})$$

Let $g : \{0, 1\}^m \rightarrow \{-1, 1\}$ be defined so that $g(x) = f(\sigma)$ using the buckets in the natural way.

2

1

7

5

3

4

6

Key facts: g is an LTF in the Boolean sense, so $\text{avg Inf}^{(i)}(g) \leq O(1/\sqrt{m})$.


The distribution of $(x, x^{(i)})$ induces $\mathbb{1}(\sigma, N_\delta(\sigma))$.

$$\text{NS}_\delta(f) = \mathbb{E}[\text{avg Inf}^{(i)}(g)] \leq O(1/\sqrt{m}) = O(\sqrt{\delta})$$


Draw a permutation that only shuffles coordinates in each bucket.

Roadmap to agnostic learning

$$f : \text{Sym}_n \rightarrow \mathbb{R}$$



Define $\text{NS}_\delta(f)$, and obtain a nice Fourier representation.

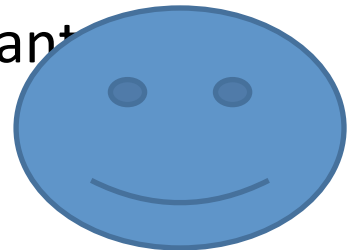


Show that the distance from $\frac{1}{2}$ for the multipliers decays at an **exponential** rate in the degree.



Bound $\text{NS}_\delta(f)$ for LTF's (and functions of LTF's by union bound).

Efficient agnostic learning over permutation invariant distributions follows from [KKMS05].



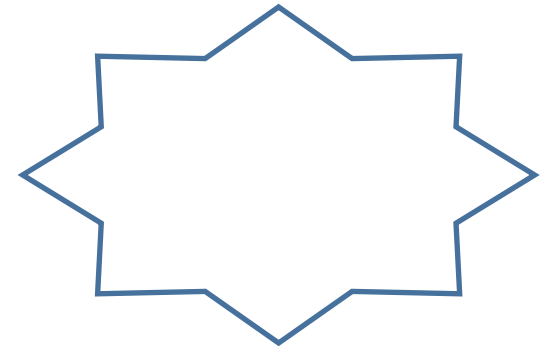
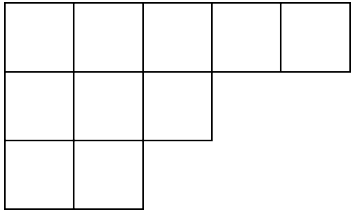
But wait, there's more!

$$\text{NS}_\delta(f) = \mathbb{E}[\text{avg Inf}^{(i)}(g)] \leq O(1/\sqrt{m}) = O(\sqrt{\delta})$$

Actually, the only LTF properties used are (a) bounded $\text{avg Inf}^{(i)}(g)$ in the Boolean analogue, and (b) closed under natural restrictions.

So the same proof works for AC^0 circuits, degree-d PTFs, etc.

Further, the time and sample complexity for permutation Invariant distributions is comparable to the uniform distribution case.



Thank you!

