Agnostically learning under permutation invariant distributions

Karl Wimmer
Mathematics & Computer Science Department
Duquesne University
Pittsburgh, PA
wimmerk@duq.edu

***Abstract*—We generalize algorithms from computational learning theory that are successful under the uniform distribution on the Boolean hypercube $\{0,1\}^n$ to algorithms successful on permutation invariant distributions. A permutation invariant distribution is a distribution where the probability mass remains constant upon permutations in the instances. While the tools in our generalization mimic those used for the Boolean hypercube, the fact that permutation invariant distributions are not product distributions presents a significant obstacle.**

Under the uniform distribution, halfspaces can be agnostically learned in polynomial time for constant ϵ. The main tools used are a theorem of Peres [Per04] bounding the *noise sensitivity* of a halfspace, a result of [KOS04] that this theorem implies Fourier concentration, and a modification of the Low-Degree algorithm of Linial, Mansour, Nisan [LMN93] made by Kalai et. al. [KKMS08]. These results are extended to arbitrary product distributions in [BOW08].

We prove analogous results for permutation invariant distributions; more generally, we work in the domain of the symmetric group. We define noise sensitivity in this setting, and show that noise sensitivity has a nice combinatorial interpretation in terms of Young tableaux. The main technical innovations involve techniques from the representation theory of the symmetric group, especially the combinatorics of Young tableaux. We show that low noise sensitivity implies concentration on "simple" components of the Fourier spectrum, and that this fact will allow us to agnostically learn halfspaces under permutation invariant distributions to constant accuracy in roughly the same time as in the uniform distribution over the Boolean hypercube case.

I. Introduction

In this paper we:

- Generalize the Low-Degree algorithm (and the agnostic learning algorithm of [KKMS08]) to the symmetric group, taking special care to account for the fact that the Fourier coefficients are matrices.
- Generalize the concept of noise sensitivity to a function $f : S_n \to \mathbb{R}$, and give an expression for noise sensitivity in terms of the Fourier spectrum of such functions. This expression will be useful for learning applications.
- Prove that the noise sensitivity of generalized linear threshold functions $f : S_n \to \mathbb{R}$ have bounded noise sensitivity.

Our primary motivation is the class of binary classification problems over the instance space $\mathcal{X} = X^n$ for some set $|X|$ with $|X| = \text{poly}(n)$. Consider the following algorithm for learning in such a scenario:

Given m examples of training data $(\overrightarrow{x}_1, y_1), \ldots, (\overrightarrow{x}_m, y_m) \in X^n \times \{-1, 1\}$,

1) Convert each instance $\overrightarrow{x}_i$ into a vector in $\{0,1\}^{n \cdot |X|}$, one bit for each attribute-value pair.
2) Consider "features" which are products of up to d of the new 0-1 attributes.
3) Find the linear function W in the feature space that best fits the training labels under some loss measure ℓ: e.g., squared loss, hinge loss, or L_1 loss.
4) Output the hypothesis $\text{sgn}(W - \theta)$, where $\theta \in [-1, 1]$ is chosen to minimize the hypothesis' training error.

We note that the above algorithm, which we will refer to as polynomial regression (as in [BOW08]), is a version of the wildly popular SVM algorithm. It is known that the above algorithm runs in $\mathrm{poly}(m, n^d)$ time, given m examples. Given $m = \Omega(n^d/\epsilon)$ examples, the SVM algorithm does generalize to unseen data, even in the case that $\mathcal{X}$ is an arbitrary product distribution [BOW08]. In fact, in [BOW08], many $\mathcal{X}$ can be the product of different sets, which can be of any finite cardinality (and even uncountably infinite under distributional assumptions). Further, to achieve any provable guarantee that the hypothesis generalizes, $m = \theta(n^d)/\epsilon$ examples are necessary.

In the literature, much effort has been put into the case where the attributes are mutually independent. In this case, the data is drawn i.i.d. from a product distribution over $\mathcal{X}$. The uniform distribution over $\{0,1\}^n$ has received much attention in a number of different scenarios [KM93], [Jac95], [BBL98], [KOS04], [MOS04], [OS07], [KKMS08], [OW09]; some of these results extend to product distributions. This was the explicit motivation of [BOW08]. In [KST09], the authors show that learning of some natural concepts is possible under a randomly chosen product distribution with high probability. In practice, the assumption that the attributes are all mutually independent is unrealistic.

The assumption we make in this work is that the distribution is *permutation invariant*. By this, we mean that $\mathbf{Pr}_{\boldsymbol{X}}[\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)] = \mathbf{Pr}_{\boldsymbol{X}}[\boldsymbol{X} = (\boldsymbol{x}_{\sigma(1)}, \boldsymbol{x}_{\sigma(2)}, \ldots, \boldsymbol{x}_{\sigma(n)})]$ for any permutation σ. A permutation invariant distribution need not even be pairwise independent, although product distributions with equal biases for each coordinate are permutation invariant. Considering this type of distribution has already proven helpful in other work. In the context of learning monotone functions over the uniform distribution on $\{0,1\}^n$, [OW09] show that product distributions are too spread out for a natural approach to work. Converting to permutation invariant distributions of the previously mentioned form allows for stronger analysis. The approach in [OW09] is to learn monotone functions over one level of the Boolean cube at a time, focusing on the instances with a fixed number of 1's at a time.

To prove our results, we use ideas from representation theory over the symmetric group. We define noise sensitivity for the symmetric group, and achieve similar learning results to those over the uniform distribution. The results and connections between computational learning theory and representation theory are interesting, and the representation theory themed results we show may be of independent interest.

A. The learning framework

Our goal is binary classification learning in the "agnostic" model introduced in [KSS94]. In this model, there is an unknown target function $t : \mathcal{X} \to \{-1, 1\}$ which we are trying to recover. Our access to t is limited; we receive labeled examples of the form $\{x, t(x)\}$, where the marginal distribution on the first coordinate is some distribution $\mathcal{D}$ on the set $\mathcal{X}$. Further, the examples are generated independently. The algorithm's task is to output a hypothesis $h : \mathcal{X} \to \{-1, 1\}$ with minimal classification error with respect to t; that is, we wish to choose h minimizing $\mathrm{err}_t(h) = \mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{D}}[h(\boldsymbol{x}) \neq t(\boldsymbol{x})]$. We compare the error of our hypothesis to the minimum error achievable using a function from some fixed class $\mathcal{C}$ of functions $\mathcal{X} \to \{-1, 1\}$. We say that we can "agnostically learn with respect to $\mathcal{C}$ under the distribution $\mathcal{D}$" if, for any target function t, any $\epsilon > 0$, given labeled examples, our algorithm returns a hypothesis satisfying

$$\mathbf{E}[\mathrm{err}_t(h)] \leq \inf_{f \in C} err_t(f) + \epsilon,$$

where the expectation is over the randomness of our algorithm.

In light of strong computational hardness results for such problems when the distribution is arbitrary, much work has gone into the case where the distribution $\mathcal{D}$ is a product distribution; that is, all of the coordinates are mutually independent. The advantage of such a distribution is the ease with which an orthogonal decomposition can be constructed and analyzed.

II. Algorithms for learning

We recount some algorithms for learning. First, we make a definition:

Definition II.1. *For a function $f : \{0,1\}^N \to \{-1,1\}$, we say that f is ϵ-concentrated to degree d with respect to $\mathcal{D}$ if there exists a polynomial of degree at most d such that $\mathbf{E}_{\boldsymbol{x}\sim\mathcal{D}}[(f(\boldsymbol{x})-p(\boldsymbol{x}))^2] \leq \epsilon$. We say that a class of functions $\mathcal{C}$ is ϵ-concentrated to degree d if every $f \in \mathcal{C}$ is ϵ-concentrated to degree d.*

The concept of ϵ-concentration has proved fruitful in computational learning theory. The Low-Degree Algorithm of Linial, Mansour, and Nisan was the first result to use ϵ-concentration. In their paper, they considered learning under the uniform distribution on $\{0,1\}^n$. They show that if the target function t is computable by a size s, depth c circuit, then t is ϵ-concentrated to degree $(O(\log(s/\epsilon)))^c\}$. We note that the original result does not hold in the agnostic framework; the assumption is that the t is in the concept class $\mathcal{C}$.

In [KOS04], Klivans, O'Donnell, and Servedio develop the "noise sensitivity method" along these lines. They show ϵ-concentration results for functions with bounded noise sensitivity. Specifically, they show that any function $f : \{0,1\}^n \to \{-1,1\}$ that can be written as a function of k linear threshold functions is ϵ-concentrated to degree $O(k^2/\epsilon^2)$ under the uniform distribution on $\{0,1\}^n$. Again, the result in their paper is only applicable to learning when t is in the concept class $\mathcal{C}$.

A significant step forward is given in [KKMS08]. In their paper, they consider the agnostic problem, where we do not assume that $t \in \mathcal{C}$. Specifically, they show that L_2-approximability bounds can be used to imply L_1-approximability bounds, which can be used to achieve results in the agnostic setting. Under the uniform distribution on $\{0,1\}^n$, they show that ϵ^2-concentration on the set $\mathcal{S}$ implies agnostic learning with accuracy loss ϵ. Their algorithm is slightly different than the Low-Degree algorithm; their algorithm solves a linear program for estimates for the Fourier coefficients indexed by $\mathcal{S}$ such that the L_1-error of the estimates is minimized. They also note that this is a version of the wildly popular Support Vector Machine (SVM) algorithm.

A further step in the distribution is given in [BOW08]. Blais, O'Donnell, and Wimmer show that nearly all results about ϵ-concentration with respect to the uniform distribution on $\{0,1\}^n$ can be applied to arbitrary product distributions; distributions where the coordinates are mutually independent, but each coordinate has an arbitrary distribution. The main tools therein are an extension of the "noise sensitivity method" of [KOS04] to product distributions and an application of the algorithm of [KKMS08]. In this paper, we use both of these techniques, suitably adjusted for functions over the symmetric group.

The main technique we will use comes from representation theory of the symmetric group. We recall the work of Boneh [Bon95], which used representation theory of groups of size 2^n to establish learning results for the uniform distribution on the hypercube. The symmetric group is not mentioned in Boneh's paper. We note that learning results of a more applied nature are known for the symmetric group, particularly in the realm of multi-object tracking [Kon08], [KHJ07], [KB08], [HGG09]. All of these results make heavy use of representation theory; this document combines the techniques used in these works with the tools and ideas from the field of computational learning theory. Further, [SJ06] introduces a similar idea which they call permutation-invariant SVMs, where the classifier is forced to be invariant under permutations. We note that our assumption is that the distribution is invariant under permutations, but the classifier need not be.

III. Overview of results

The main result in this paper concerns learning linear threshold functions, which we define here.

Definition III.1. *We say that $f : X^n \to \{-1,1\}$ is a* linear threshold function *if its "attribute-value pair" analogue $f : \{0,1\}^{n\cdot|X|}$ is a linear threshold function.*

Each attribute-value pair is a $\{0,1\}$ indicator of the event $X_i = x_j$, where $1 \leq i \leq n$ and $1 \leq j \leq |X|$. We will also consider linear threshold functions $f : S_n \to \{-1,1\}$, which we massage into the above definition by encoding a permutation as n^2 attribute-value pairs: the indicators of $\sigma(i) = j$ for $1 \leq i,j \leq n$.

Theorem III.2. *Let $\mathcal{C}$ be the class of functions of k linear threshold functions over $\{0,1,2,\ldots,B-$*

$1\}^n$*, and let* $\mathcal{D}$ *be any permutation invariant distribution over* $\{0, 1, 2, \ldots, B-1\}^n$*. There is an algorithm that agnostically learns with respect to* $\mathcal{C}$ *under the distribution* $\mathcal{D}$*, using* $n^{O(k^2/\epsilon^4+B)}$ *time and examples.*

We remark that the attribute-value pair distribution is not a permutation invariant distribution.

Corollary III.3. *Let* $\mathcal{C}$ *be the class of functions of* k *linear threshold functions over* $\{0,1\}^n$*, and let* $\mathcal{D}$ *be any permutation invariant distribution over* $\{0,1\}^n$*. There is an algorithm that agnostically learns with respect to* $\mathcal{C}$ *under the distribution* $\mathcal{D}$*, using* $n^{O(k^2/\epsilon^4)}$ *time and examples.*

We note that in a permutation invariant distribution, the attributes need not even be pairwise independent. The class of permutation invariant distributions includes some product distributions as a special case. Specifically, any product distribution where all the bits have the same bias is a permutation invariant distribution. In some sense, permutation invariant distributions are a generalization of p-biased distributions.

To prove Theorem III.2, we will prove the following theorem:

Theorem III.4. *Let* $\mathcal{C}$ *be the class of functions of* k *linear threshold functions over* S_n*. There is an algorithm that agnostically learns* $\mathcal{C}$ *under the uniform distribution on* $\mathcal{D}$*, using* $n^{O(k^2/\epsilon^4)}$ *time and examples.*

For consistency with our previous definitions, here we take $X = \{1, 2, 3, \ldots, n\}$, and $\mathcal{D}$ is the distribution supported solely on permutations of the vector $(1, 2, 3, \ldots, n)$. We can identify a permutation of this vector with an element of $\sigma \in S_n$ by setting the ith element of the permuted vector to be $\sigma(i)$.

To prove Theorem III.4, we consider the well-studied noise sensitivity method first introduced in [KOS04]. However, it is not immediately clear how noise should be defined in the case of a nonproduct distribution; we were unable to find a definition in the literature. We give an informal definition here.

Imagine that Alice wishes to communicate a permutation $\sigma \in S_n$ to Bob. She will attempt to do so by transmitting $\sigma(1)$, then $\sigma(2)$, and so on until $\sigma(n)$. However, each transmission has a probability of $1-\delta$ of succeeding in reaching Bob; for each $i \in [n]$, with probability δ no information about $\sigma(i)$ is relayed to Bob. Bob selects a permutation σ' uniformly at random from the set of permutations consistent with the information he received from Alice. With respect to some function $f : S_n \to \{-1, 1\}$, we can define the noise sensitivity of f to be $\mathbb{NS}_\delta(f) = \mathbf{Pr}_{\boldsymbol{\sigma},\boldsymbol{\sigma}'}[f(\boldsymbol{\sigma}) \neq f(\boldsymbol{\sigma}')]$ where $\boldsymbol{\sigma}$ is uniformly distributed over S_n, and $\boldsymbol{\sigma}'$ is a random permutation consistent with the successfully received images of $\boldsymbol{\sigma}$.

We first formalize our definition of noise sensitivity in Section IV. In Section V, we introduce machinery from representation theory. We tie noise sensitivity and representation theory together and sketch the technical theorems needed to establish that bounded noise sensitivity implies ϵ-concentration in Section VI. We generalize a theorem of Peres [Per04], obtaining a useful bound the noise sensitivity of linear threshold functions in the domain of the symmetric group in Section VII; this step does not require any representation theory. Finally, in section VIII, we prove Theorem III.2 from Theorem III.4 in a manner similar to [BOW08].

IV. Noise sensitivity

To formalize our noise sensitivity experiment, we make an intermediate definition.

Definition IV.1. *With respect to* S_n*, let* N^q *be the uniform distribution over all permutations with at least* $n-q$ *fixed points.*

In our previously mentioned scenario with Alice and Bob, the fixed points are the images $\sigma(i)$ successfully received. Our actual noise experiment will be a mixture of N^q distributions, where we will choose q randomly.

Definition IV.2. *With respect to* S_n*, let* N_δ *be the distribution whose samples are generated in the following way: let* $\boldsymbol{q}$ *be a binomially distributed random variable with* n *trials and a success probability of* δ*. Then* N_δ *is a random draw from* $N^{\boldsymbol{q}}$*.*

It what follows, $\boldsymbol{q}$ is always a binomially distributed random variable with n trials and success probability δ.

Definition IV.3. *Let $N^q(\sigma)$ ($N_\delta(\sigma)$) be the distribution $\psi\sigma$, where ψ is chosen from N^q (N_δ).*

The set of permutations having at most q fixed points is a union of conjugacy classes, so N^q is a class distribution, meaning N^q is uniform on conjugacy classes. It follows that if ψ is chosen from N^q, then $\psi\sigma$ and $\sigma\psi$ have the same distribution for every σ. Also, N_δ is a class distribution, because it is a mixture of class distributions. In our definition of $N^q(\sigma)$, we could have taken $\sigma\psi$ instead of $\psi\sigma$.

Definition IV.4. *We define the functional operator T_δ such that, given a function $f : S_n \to \mathbb{R}$.*

$$(T_\delta f)(\sigma) = \mathop{\mathbf{E}}_{\psi\sim N_\delta(\sigma)}[f(\psi)] = \mathop{\mathbf{E}}_{\psi\sim N_\delta}[f(\psi\sigma)] = \mathop{\mathbf{E}}_{\psi\sim N_\delta}[f(\sigma\psi)].$$

Definition IV.5. *If $f : S_n \to \{-1, 1\}$, then*

$$\mathbb{NS}_\delta(f) = \mathop{\mathbf{Pr}}_{\boldsymbol{\sigma},\psi\sim N_\delta(\boldsymbol{\sigma})}[f(\boldsymbol{\sigma}) \neq f(\psi)]$$

It is not hard to show that $\mathbb{NS}_\delta(f) = \frac{1}{2} - \frac{1}{2}\mathbf{E}_{\boldsymbol{\sigma},\psi\sim N_\delta(\boldsymbol{\sigma})}[f(\boldsymbol{\sigma})f(\psi)] = \frac{1}{2} - \frac{1}{2}\langle f, T_\delta(f)\rangle$, where we have implicitly defined the inner product $\langle f, g\rangle = \mathbf{E}_{\boldsymbol{\sigma}}[f(\boldsymbol{\sigma})g(\boldsymbol{\sigma})]$.

We prove the following generalization of Peres' Theorem [Per04]:

Theorem IV.6. *Let $f : S_n \to \{-1, 1\}$ be a linear threshold function. Then $\mathbb{NS}_\delta(f) \leq O(\sqrt{\delta})$.*

We note that the bound $O(\sqrt{\delta})$ is the same bound for product distribution versions of this theorem, and the constant is no larger than the constant for the uniform distribution. The advancement of noise sensitivity bounds is very recent; the uniform distribution case is discussed in [Per04], and the arbitrary product distribution case is proved in [BOW08].

V. Representation theory

Ideally, noise sensitivity should have a nice interpretation in terms of the Fourier spectrum of f, as in the case of functions whose domain is $\{0, 1\}^n$. It is not clear that such an interpretation should exist, much less that it should be useful. When we appeal to the Fourier spectrum of a function $f : S_n \to \mathbb{R}$, the components of the Fourier spectrum are matrices, not all scalars. However, we show that with a suitable definition of noise, noise sensitivity is well-behaved; the noise sensitivity involves multiplying each component of the Fourier spectrum by some scalar independent of f. Finally, we show that we can use our interpretation of noise sensitivity along with our noise sensitivity bound to achieve ϵ-concentration. Applying the result of [KKMS08] finishes the claim.

To use tools from Fourier analysis, we first mention background about representation theory of finite groups.

Definition V.1. *We say that a* representation *ρ of a group G is a mapping $\rho : G \to \mathbb{R}^{d_\rho \times d_\rho}$ which preserves the algebraic structure of G; that is, for $\sigma_1, \sigma_2 \in G$ we have $\rho(\sigma_1\sigma_2) = \rho(\sigma_1) \cdot \rho(\sigma_2)$. The matrices in the codomain of ρ are called the representation matrices, and d_ρ is the degree of the representation.*

We will be concerned with representations that are *irreducible*. These are representations that can not be decomposed into simpler representations (for some suitable definition of decompose). A set of irreducible representations contains all information about the structure of G. We will not be concerned with any specific set of irreducible representations, but we mention that there are certain canonical choices. As an aside, if G is abelian, then all the irreducible representations have degree 1.

The Peter-Weyl Theorem says that the functions given in the matrix entries of irreducible representations of G form an orthogonal basis for $L^2(G)$. For any group G, we have the following definition.

Definition V.2. *Let $f : G \to \mathbb{R}$ be any function on G, and let ρ be any representation on G. The Fourier coefficient of f at the representation ρ is given by the matrix*

$$\hat{f}_\rho = \frac{1}{|G|}\sum_{\sigma\in G} f(\sigma)\rho(\sigma)$$

The collection of the matrices $\hat{f}_\rho$ at irreducible representations of G is called the Fourier transform of f.

To reconstruct f from its Fourier transform, we need the Fourier Inversion formula:

$$f(\sigma) = \sum_{\rho} d_\rho \mathrm{tr}\left(\hat{f}_\rho^T \rho(\sigma)\right).$$

The tr expression can be thought of as a dot product between two length-d_ρ^2 vector versions of $\hat{f}_\rho$ and ρ, arranging each matrix into a vector by taking the elements first top to bottom, then left to right as vectors.

As an analogue to sum of squares of Fourier coefficients, we will use the Frobenius norm of a matrix. If A is a square n-by-n matrix and contains only real valued entries, then $\|A\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij})^2}$. We also have results such as Parseval's theorem in this setting; specifically, $\sum_{\rho} d_\rho \|\hat{f}_\rho\|^2 = 1$ for functions whose range is $\{-1, 1\}$.

A. The symmetric group and Young tableaux

We will be very interested in the irreducible representations of S_n. We define a partition λ of n to be a nonincreasing sequence of integers $(\lambda_1, \lambda_2, \ldots, \lambda_m)$, where $\sum_i \lambda_i = n$ and each $\lambda_i > 0$. We write this as $\lambda \vdash n$. The following is well-known:

Theorem V.3. *The irreducible representations of S_n are indexed by partitions of n.*

It is common to use Ferrer's diagrams to visualize a partition of n. The Ferrer's diagrams represent each component of each partition as the number of squares in the corresponding row. A standard Young tableaux is a Ferrer's diagrams with the numbers $1, 2, \ldots, n$ each occurring in one cell, such that the numbers in the cells are increasing downwards and to the right. (We refer the reader to [Ful97] for a thorough treatment of Young tableaux.) We define the dominance order on partitions of n in the following way: we write $\lambda \trianglerighteq \mu$ if $\sum_{i=1}^{k} \lambda_i \geq \sum_{i=1}^{k} \mu_i$ for all k, where we pad the partitions with extra zeroes.

Another combinatorial concept that will be important for us is the concept of skew diagrams. Given a partition $\lambda \vdash n$, and $\mu \vdash k$, a skew diagram of shape λ/μ is the diagram formed by the set-theoretic difference of λ and μ. A skew-standard Young tableau is a skew diagram filled with 1 up to the number of cells in the diagram (which may be more than $n-k$), each number occurring in exactly one cell, where the numbers increase downwards and to the right. We will only be concerned with skew tableaux with $\mu = (k)$.

1	2	5
3	6	
4		

A standard Young tableau of shape (3,2,1)

	1	4	5
2	3		

A skew-standard Young tableau of shape (4,2)/(1)

We say that the *degree* of a Ferrer's diagram of shape λ is the number of standard Young tableaux of shape λ, and we denote the degree $\dim \lambda$ or d_λ. We will also refer to this as the degree of the partition λ. Similarly, we define $\dim \lambda/\mu$ as the number of skew-standard Young tableaux of shape λ/μ. A variety of expressions for these dimensions are known.

Because of the equivalence of irreducible representations and partitions, we will frequently identify ρ_λ and λ, where ρ_λ is an irreducible representation corresponding to the class of representations indexed by the partition λ. We will shorten ρ_λ to ρ when the correspondence is clear from context. The fact that we called the degree of a Ferrer's diagram d_λ is suggestive of the following theorem:

Theorem V.4. *The degree of any irreducible representation is equal to the degree of the Ferrer's diagram of the corresponding partition. That is, $d_\lambda = d_{\rho_\lambda}$, which we will write as d_ρ when clear from context. Further, we can write $\widehat{f_\lambda}$ in place of $\widehat{f_\rho}$.*

We remark here that it is well-known (see, for example, [HGG09]) that the representations corresponding to partitions λ where $\lambda_1 > n - k$ span the vector space spanned by all functions that depend on k coordinates. This will allow us to use the polynomial regression algorithm stated near the beginning at the document. More specifically, using Plancherel's theorem, we can satisfy the definition of ϵ-concentration up to degree $d-1$ by showing

$$\sum_{\lambda:\lambda_1 \leq n-d} d_\lambda \|\hat{f}_\lambda\|^2 \leq \epsilon.$$

The associated polynomial can be taken to be equivalent to $\sum_{\lambda:\lambda_1 > n-d} \hat{f}_\lambda$ when the domains are suitably adjusted.

VI. Sketch of technical results

The proofs of the theorems in this section are omitted due to space restrictions.

Theorem VI.1. *Given $f : S_n \to \mathbb{R}$ and $\delta > 0$, there exist constants $c_{\lambda,\delta}$ independent of f for each $\lambda \vdash n$ such that*

$$\langle f, T_\delta f \rangle = \sum_{\lambda \vdash n} c_{\lambda,\delta} d_\lambda \|\widehat{f}_\lambda\|^2.$$

This result tells us that noise is independent of the irreducible representations used, and only depends on the corresponding partition λ indexing the representation. Using this result, we get:

$$\begin{aligned} \mathbb{NS}_\delta(f) &= \tfrac{1}{2} - \tfrac{1}{2}\langle f, T_\delta f \rangle \\ &= \tfrac{1}{2} - \tfrac{1}{2}\sum_{\lambda \vdash n} c_{\lambda,\delta} d_\lambda \|\hat{f}_\lambda\|^2 \end{aligned}$$

Restricting f to be $\{-1,1\}$-valued and assuming $\mathbb{NS}_\delta(f) \leq \epsilon$, we have:

$$\begin{aligned} \epsilon &\geq \tfrac{1}{2} - \tfrac{1}{2}\sum_{\lambda \vdash n} c_{\lambda,\delta} d_\lambda \|\hat{f}_\lambda\|^2 \\ &= \tfrac{1}{2}\Big(\sum_{\lambda \vdash n} d_\lambda \|\hat{f}_\lambda\|^2 - \sum_{\lambda \vdash n} c_{\lambda,\delta} d_\lambda \|\hat{f}_\lambda\|^2\Big) \\ &= \tfrac{1}{2}\Big(\sum_{\lambda \vdash n} (1 - c_{\lambda,\delta}) d_\lambda \|\hat{f}_\lambda\|^2\Big) \\ &\geq \tfrac{1}{2}\Big(\sum_{\lambda:\lambda_1 \leq n-1/\delta} (1 - c_{\lambda,\delta}) d_\lambda \|\hat{f}_\lambda\|^2\Big), \end{aligned}$$

using Parseval's identity in the first equality.

We will show that functions $f : S_n \to \{-1,1\}$ satisfying $\mathbb{NS}_\delta(f) \leq \epsilon$ are $O(\epsilon)$-concentrated up to degree $1/\delta$ for by showing that

$$\max_{\lambda:\lambda_1 \leq n-1/\delta} c_{\lambda,\delta} \leq 2e^{-0.9} < 1$$

Intuitively, these coefficients $c_{\lambda,\delta}$ should decrease as λ becomes more "complex." This is indeed the case. In the Boolean hypercube case, it is fairly straightforward to determine the values of these coefficients; it is significantly more difficult in our setting. The bulk of our effort goes into analyzing these coefficients. For technical reasons, we will assume $\delta > 12/n$.

We prove a connection between our noise operator on the symmetric group and Young tableaux. This result requires analysis of group characters, which we do not discuss here. It also holds for real-valued functions on S_n.

Theorem VI.2. *Let ρ be some irreducible representation, and λ its corresponding partition. Then the Fourier spectrum of $T_\delta f$ consists of matrices of the form $\mathbf{E}_{\boldsymbol{q}}[\frac{\dim \lambda/(\boldsymbol{q})}{d_\rho}]\widehat{f}_\rho$ at the representation ρ, where the Fourier spectrum of f consists of the matrices $\widehat{f}_\rho$.*

Thus $c_{\lambda,\delta} = \mathbf{E}_{\boldsymbol{q}}[\frac{\dim \lambda/(\boldsymbol{q})}{d_\lambda}]$, where $\boldsymbol{q}$ is a binomially distributed random variable with n trials and success probability δ.

We turn our attention to expressions of the form $\frac{\dim \lambda/(q)}{d_\lambda}$. Intuitively, these expectations should decrease as λ (and ρ) become more complex. We use a deep combinatorial result of [OO96]:

$$\frac{\dim \lambda/(q)}{d_\lambda} = \frac{(n-q)!}{n!} \sum_{i_1 \geq i_2 \geq \ldots \geq i_q \geq 1} \prod_{j=1}^{q} (\lambda_{i_j} - j + 1)$$

Using this result, we prove the following theorem, which may be of independent interest:

Theorem VI.3. *Let $\lambda \vdash n, \beta \vdash n$, and $\beta \trianglelefteq \lambda$ in the dominance ordering. Then for any integer $q \geq 0$,*

$$\frac{\dim \beta/(q)}{d_\beta} \leq \frac{\dim \lambda/(q)}{d_\lambda}.$$

It follows that $\max_{\lambda:\lambda_1 \leq n-1/\delta} c_{\lambda,\delta}$ occurs when $\lambda = (n - 1/\delta, 1/\delta)$, so it suffices to consider partitions into two parts.

Theorem VI.4. *Let $\lambda \vdash n$, where $\lambda = (\lambda_1, \lambda_2)$. Then*

$$\frac{\dim \lambda/(q)}{d_\lambda} \leq \left(\frac{\dim \lambda/(2)}{d_\lambda}\right)^{q-1}$$

Set $\lambda = (n - 1/\delta, 1/\delta)$. We analyze $c_{\lambda,\delta}$ as follows:

$$\mathbf{E}_{\boldsymbol{q}}[\frac{\dim \lambda/(\boldsymbol{q})}{d_\lambda}] \leq \mathbf{E}_{\boldsymbol{q}}[\left(\frac{\dim \lambda/(2)}{d_\lambda}\right)^{\boldsymbol{q}-1}]$$
$$= \frac{d_\lambda}{\dim \lambda/(2)} \mathbf{E}_{\boldsymbol{q}}[\left(\frac{\dim \lambda/(2)}{d_\lambda}\right)^{\boldsymbol{q}}]$$
$$= \frac{d_\lambda}{\dim \lambda/(2)} \left(1 - \left(1 - \frac{\dim \lambda/(2)}{d_\lambda}\right)\delta\right)^n. \quad (1)$$

Using Equation VI, it is not difficult to show that $\frac{\dim \lambda/(2)}{d_\lambda} = 1 - \frac{(1/\delta)(n - 1/\delta + 1)}{n(n-1)}$. Also, $\frac{d_\lambda}{\dim \lambda/(2)} \leq 2$, since $1/\delta \geq n/2$.

Continuing from (1), we have

$$\frac{d_\lambda}{\dim \lambda/(2)} \left(1 - \left(1 - \frac{\dim \lambda/(2)}{d_\lambda}\right)\delta\right)^n \leq$$
$$\frac{d_\lambda}{\dim \lambda/(2)} \left(1 - \frac{(n - 1/\delta + 1)}{n(n-1)}\right)^n \leq$$
$$2\exp(-\frac{(n - 1/\delta + 1)}{(n-1)}) \leq 2\exp(-0.9)$$

when $\delta > 12/n$. Therefore, $\max_{\lambda:\lambda_1 \leq n - 1/\delta} c_{\lambda,\delta} \leq 2e^{-0.9}$, and

$$\sum_{\lambda:\lambda_1 \leq n-1/\delta} d_\lambda \|\hat{f}_\lambda\|^2 \leq \frac{2}{1 - 2e^{-0.9}}\epsilon.$$

Given a bound on noise sensitivity of the form $\mathbb{NS}_\delta(f) \leq g(\delta)$ for some function g decreasing to 0 as δ goes to 0, we can achieve ϵ^2-concentration by choosing a sufficiently small value for δ. This is sufficient to invoke the polynomial regression algorithm along the lines of [KKMS08] for agnostic learning of functions with low noise sensitivity.

VII. Bounding noise sensitivity

We remind the reader of our definition for linear threshold functions $f : S_n \to \{-1, 1\}$, which is more convenient for our purposes here.

Definition VII.1. *We say that a function $f : S_n \to \{-1, 1\}$ is a linear threshold function if its "attribute-value pair" function $\{0, 1\}^{n^2} \to \{-1, 1\}$ is a linear threshold function.*

Another way to think of this is to encode permutations using their permutation matrices in $\{0, 1\}^{n \times n}$, then converting to a vector in $\{0, 1\}^{n^2}$ by taking the columns of such a permutation matrix in column major order.

Theorem VII.2. *Let $f : S_n \to \{-1, 1\}$ be a linear threshold function. Then $\mathbb{NS}_\delta(f) \leq 2\sqrt{\delta}$, under the uniform distribution on S_n.*

Proof: Our proof will closely mirror the proof of Peres' Theorem given in [O'D07]. As in [O'D07], we prove a slightly stronger statement. Let $F_1, \ldots, F_m$ be any partition of n. For each F_i, we have an associated permutation ψ_i such that every point of $[n] - F_i$ is a fixed point of ψ_i. We then apply one of the ψ_i's to σ, chosen uniformly at random.

The first statement we prove is the following:

Lemma VII.3. *Let $f : S_n \to \{-1, 1\}$ be a linear threshold function, let $F_1, \ldots, F_m \subseteq [n]$ be a partition of n, and let $\psi_1, \ldots, \psi_m$ be permutations as previously mentioned. Then $\mathbf{Pr}_{\boldsymbol{\sigma},\boldsymbol{i}}[f(\boldsymbol{\sigma}) \neq f(\psi_{\boldsymbol{i}}\boldsymbol{\sigma})] \leq m^{-1/2}$.*

When $m = 1/\delta$, it is straightforward to check that the distribution on $\boldsymbol{\psi}_i\sigma$ is the same as our previously mentioned noise experiment, when the partition and the ψ_i's are chosen uniformly at random as well. Every coordinate in σ is non-fixed with probability δ.

Proof: We note that all the ψ_i's commute. We identify every string $x \in \{0, 1\}^m$ with the permutation $\psi_{I_1(x)}\boldsymbol{\sigma}$, where $\psi_{I_1(x)}$ denotes the composition of the ψ_i's where $x_i = 1$. We note that $\psi_{I_1(x)}\boldsymbol{\sigma}$ is uniformly distributed on S_n if $\boldsymbol{\sigma}$ is, for any x. We define $g(x) = f(\psi_{I_1(x)}\boldsymbol{\sigma})$. We have

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\sigma},\boldsymbol{i}}[f(\boldsymbol{\sigma}) \neq f(\psi_{\boldsymbol{i}}\boldsymbol{\sigma})] = \mathop{\mathbf{E}}_{\boldsymbol{\sigma}}[\mathop{\mathbf{Pr}}_{\boldsymbol{i}}[f(\boldsymbol{\sigma}) \neq f(\psi_{\boldsymbol{i}}\boldsymbol{\sigma})]]$$
$$= \mathop{\mathbf{E}}_{\boldsymbol{\sigma},\boldsymbol{x}}[\mathop{\mathbf{Pr}}_{\boldsymbol{i}}[f(\psi_{I_1(\boldsymbol{x})}\boldsymbol{\sigma}) \neq f(\psi_{I_1(\boldsymbol{x}^{(i)})}\boldsymbol{\sigma})]]$$
$$= \mathop{\mathbf{E}}_{\boldsymbol{\sigma},\boldsymbol{x}}[\mathop{\mathbf{Pr}}_{\boldsymbol{i}}[g(\boldsymbol{x}) \neq g(\boldsymbol{x}^{(i)})]].$$

Consider $\mathbf{Pr}_{\boldsymbol{x}}[g(\boldsymbol{x}) \neq g(\boldsymbol{x}^{(i)})]$, where $\boldsymbol{x}^{(i)}$ denotes the string $\boldsymbol{x}$ with the ith bit flipped. Since f is a linear threshold function and the ϕ_i's are all disjoint, g is also a linear threshold function (in the traditional binary sense). Further, $\mathbf{Pr}_{\boldsymbol{i}}[g(\boldsymbol{x}) \neq g(\boldsymbol{x}^{(i)})]$ is the average influence of g, which is known to be at most $m^{-1/2}$ when g is a linear

threshold function. This completes the proof of the lemma. ∎

To complete the proof, we notice that the noise experiment can be thought of in the following way, when $\delta = 1/m$ for an integer m: For each $i \in [n]$, put i in one of $F_1, \ldots, F_m$ uniformly at random. Every coordinate has an $1/m = \delta$ chance of being in the set F_i that $\boldsymbol{\psi}_i$ selects its potential non-fixed points, and a random permutation is chosen restricted to those coordinates.

Using the lemma:

$$\mathbb{NS}_\delta(f) = \mathop{\mathbf{E}}_{\boldsymbol{F}_1,\ldots,\boldsymbol{F}_m,\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_m}[\mathop{\mathbf{Pr}}_{\boldsymbol{\sigma},\boldsymbol{i}}[f(\boldsymbol{\sigma}) \neq f(\boldsymbol{\psi}_i\boldsymbol{\sigma})]] \leq m^{-1/2} \leq \sqrt{\delta}$$

If δ is not the reciprocal of an integer, we use the fact that noise sensitivity is increasing in δ, and round up to the nearest reciprocal of an integer, which can not cause us to pay more a factor of 2, increasing our bound to $2\sqrt{\delta}$. ∎

Similar to [BOW08], this proof transfers to any class of functions that is closed under complementation and restriction and has an upper bound on average influence over the uniform distribution on the Boolean hypercube. For example, it is possible to bound noise sensitivity of (suitably-defined) AC^0 circuits over the symmetric group. Analogously to the symmetric group case, we define our AC^0 circuits to take as inputs the entries of the "attribute-value" pair encoding. We leave full details to the reader.

VIII. Learning applications

For a string $x \in \{0,1\}^n$, define $I_1(x)$ as the set of indices where $x_i = 1$. Define $I_0(x)$ similarly. Define $\mathcal{U}_m$ to be the uniform distribution over strings where $|I_1(x)| = m$.

Theorem VIII.1. *There is an algorithm running in time $n^{O(1/\epsilon^4)}$ for agnostically learning with respect to the class of halfspaces over $\{0,1\}^n$ under the distribution $\mathcal{U}_m$.*

Proof: Let $f^*(x) = \operatorname{sgn}(\sum_i w_i x_i)$ be the most accurate linear threshold function in computing the target function t. From the previous section, we know $\mathbb{NS}_{1/\epsilon^2}(f^*(x)) \leq O(\epsilon)$, so $f^*(x)$ is $O(\epsilon)$-concentrated on functions up to order $1/\epsilon^2$.

Convert every example $\langle x, f(x)\rangle$ where x is drawn from $\mathcal{U}_m$ to a permutation σ by uniformly randomly assigning a random permutation from $I_1(x) \to [k]$ and $I_0(x) \to [n] \setminus [k]$. Note that there exists a linear threshold function over permutations that classifies at least as well as f^*, since the classifier

$$g(\sigma) = \operatorname{sgn}(\sum_i w_i \mathbf{1}[\sigma(i) \in [k]])$$

is a linear threshold function, and is consistent with f^*. Further, every permutation is equally likely, so the resulting distribution under this transformation is the uniform distribution over S_n. Therefore, the algorithm will output a hypothesis with error at most ϵ worse than the error of g. To classify future instances, we can convert from a bit string to a permutation in the same way. ∎

Following the noise sensitivity bound for functions of k linear threshold functions given in [KOS04], we get the following:

Theorem VIII.2. *There is an algorithm running in time $n^{O(k^2/\epsilon^4)}$ for agnostically learning with respect to the class of functions of k halfspaces over $\{0,1\}^n$ under the distribution $\mathcal{U}_m$.*

One way of viewing the above learning results is that the algorithm learns under the uniform distribution over all permutations of the string 1^k0^{n-k}. This reduction can be applied to learn over all permutations of any n-character string. For example, the symmetric group case is the case where all the characters are different; we take $X = [n]$, and $\mathcal{D}$ is the distribution that is uniform over all permutations of $(1, 2, \ldots, n)$.

Theorem VIII.3. *Let $\mathcal{D}$ be any permutation invariant distribution over $\{0,1\}^n$. There is an algorithm running in time $n^{O(k^2/\epsilon^4)}$ for agnostically learning with respect to the class of functions of k halfspaces over $\{0,1\}^n$ under the distribution $\mathcal{D}$.*

Proof: Note that it is easy to achieve perfect clustering; by observing the number of ones in each string, we can tell from which distribution $\mathcal{U}_m$ any string comes from. Following the approach in [BOW08], we can partition our examples into $n+1$ bins, one for each possible $\mathcal{U}_m$ with $0 \leq m \leq n$, and learn over each bin separately once

enough examples are seen. We note that we may have to account for some loss due to a insufficient number of examples in some bin, but this can be factored into the ϵ. ■

Theorem VIII.4. *Let $\mathcal{D}$ be any permutation invariant distribution over $\{0,1,2,\ldots,B-1\}^n$ for constant B. There is an algorithm running in time $n^{O(k^2/\epsilon^4+B)}$ for agnostically learning with respect to the class of functions of k linear threshold functions over $\{0,1\}^n$ under the distribution $\mathcal{D}$.*

Proof: The proof is virtually the same as the proof of Theorem VIII.3, except that the number of bins is now $\binom{n+B-1}{B-1} = \text{poly}(n^B)$, via a "stars and bars" counting argument. ■

IX. Conclusions

We have shown a setting far removed from the product distribution assumption in which agnostic learning is achievable. Specifically, we can efficiently agnostically learn linear threshold functions (and AC^0 circuits) over permutation invariant distributions. Our main technique is the "noise sensitivity" method, coupled with Fourier analysis. We leave open many learning problems with the domain of the symmetric group as well as the question of other groups where this approach could be successful.

References

[BBL98] Avrim Blum, Carl Burch, and John Langford. On learning monotone boolean functions. In *FOCS*, pages 408–415, 1998.

[Bon95] Dan Boneh. Learning using group representations (extended abstract). In *COLT '95: Proceedings of the eighth annual conference on Computational learning theory*, pages 418–426, New York, NY, USA, 1995. ACM.

[BOW08] Eric Blais, Ryan O'Donnell, and Karl Wimmer. Polynomial regression under arbitrary product distributions. In *Proc. 21st Workshop on Comp. Learning Theory*, 2008.

[Ful97] Willam Fulton. *Young Tableaux: with Applications to Representation Theory and Geometry*. Cambridge University Press, 1997.

[HGG09] Jonathan Huang, Carlos Guestrin, and Leonidas Guibas. Fourier theoretic probabilistic inference over permutations. *J. Mach. Learn. Res.*, 10:997–1070, 2009.

[Jac95] Jeffrey Jackson. *The Harmonic Sieve: A Novel Application of Fourier Analysis to Machine Learning Theory and Practice*. PhD thesis, Carnegie Mellon University, August 1995.

[KB08] Risi Kondor and Karsten M. Borgwardt. The skew spectrum of graphs. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 496–503, New York, NY, USA, 2008. ACM.

[KHJ07] Risi Kondor, Andrew Howard, and Tony Jebara. Multi-object tracking with representations of the symmetric group. In *In AISTATS*, 2007.

[KKMS08] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KM93] Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, December 1993. Earlier version appeared in *Proceedings of the Twenty Third Annual ACM Symposium on Theory of Computing*, pages 455–464, 1991.

[Kon08] Risi Kondor. *Group Theoretical Methods in Machine Learning*. PhD thesis, Columbia University, 2008.

[KOS04] Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.*, 68(4):808–840, 2004.

[KSS94] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2–3):115–141, 1994.

[KST09] Adam Tauman Kalai, Alex Samorodnitsky, and Shang-Hua Teng. Learning and smoothed analysis. *Foundations of Computer Science, Annual IEEE Symposium on*, 0:395–404, 2009.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *Journal of the ACM*, 40(3):607–620, 1993.

[MOS04] Elchanan Mossel, Ryan O'Donnell, and Rocco A. Servedio. Learning functions of k relevant variables. *J. Comput. Syst. Sci.*, 69(3):421–434, 2004.

[O'D07] Ryan O'Donnell. Analysis of boolean functions lecture notes, 2007. http://www.cs.cmu.edu/ odonnell/boolean-analysis/.

[OO96] Andrei Okounkov and Grigori Olshanski. Shifted schur functions. *St. Petersburg Math. J*, 9:239–300, 1996.

[OS07] Ryan O'Donnell and Rocco Servedio. Learning monotone decision trees in polynomial time. *SIAM J. Comput.*, 37(3):827–844, 2007.

[OW09] Ryan O'Donnell and Karl Wimmer. KKL, Kruskal-Katona, and monotone nets. In *FOCS*, pages 725–734, 2009.

[Per04] Yuval Peres. Noise stability of weighted majority. `arXiv:math/0412377v1`, 2004.

[SJ06] Pannagadatta K. Shivaswamy and Tony Jebara. Permutation invariant SVMs. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 817–824, New York, NY, USA, 2006. ACM.