ISSN 1342-2812

Research Reports on Mathematical and Computing Sciences

Fourier Approximation under Non-uniform Distribution

Hiroki Yamaguchi

September 2011, C–275

Department of
Mathematical and
Computing Sciences
Tokyo Institute of Technology

SERIES C: Computer Science

Abstract

We propose a generalization of a PAC-learning algorithm known as the *Low Degree Learning Algorithm* for non-uniform distributions. We show that our algorithm works under a non-uniform distribution D if the smallest eigenvalue of the Fourier coefficient matrix of the distribution is not too small. We also show that this condition is guaranteed if $|S| < 2^{\eta n}$ for some $\eta < 1$ and for each size parameter n, where S is the set of example instances of size n whose probability under D is $< (\mathrm{poly}(n)2^n)^{-1}$.

1 Introduction

The learnability notion in the *Probably Approximately Correct (PAC) learning* model is one of the most important criteria in the field of computational learning theory. Unfortunately, however, since the PAC learning model was introduced by Valiant in 1984 [7], not so many positive results have been shown on the original version of the PAC learnability problem despite of various deep research investigations of a quarter century. On the other hand, the distribution restricting version of PAC leaning model, in particular, with respect to the uniform distribution, some important learning algorithms have been obtained. Thus, it has been one of the important problems to generalize such learning algorithms for arbitrary distributions. In this paper we propose a generalization of one of such algorithms known as the *Low Degree Learning Algorithm* for non-uniform distributions.

In the PAC learning model, the learner tries to learn an unknown labeling function by a small number of pairs of data and the corresponding label which is drawn according to an also unknown probability distribution. A *concept class* $\mathbf{C}$ is a set of families of functions $f = \{f_n\}_n$, where each f_n is a function mapping $\{-1,+1\}^n$ to $\{-1,+1\}$. Let $f \in \mathbf{C}$ be an unknown *target function* which we wish to learn, and $D = \{D_n\}_n$ be a family of unknown probability distributions on each example instance $\{-1,+1\}^n$. We will not display n without need to avoid confusion. A *example oracle* $\mathrm{Ex}(f,D)$ which generates, when queried, an example $(x, f(x))$ where instance x is drawn with probability $D(x)$. A learning algorithm $\mathcal{A}$ takes parameters $\epsilon, \delta > 0$. During execution, $\mathcal{A}$ has an access to $\mathrm{Ex}(f,D)$, which is only the way for $\mathcal{A}$ to get information of f and D. The goal of $\mathcal{A}$ is, for all n, to output a hypothesis $h : \{-1,+1\}^n \to \{-1,+1\}$ such that

$$\Pr_{x \sim D}[f(x) \neq h(x)] \leq \epsilon \tag{1}$$

with probability not less than $1 - \delta$. We say that the concept class $\mathbf{C}$ is *PAC learnable* if such an algorithm $\mathcal{A}$ exists. For many basic concept classes such as DNFs, $\mathbf{AC}^0$ circuits, etc., the problem of their PAC learnability has been an important open problem.

Some notable progress has been made for the PAC learnability under the *uniform* distribution, and several important learning algorithms have been shown [1, 2, 3] for this restricted version of PAC learnability. The Fourier analysis is a fundamental tool for these results. As we will see in Section 2, the Fourier basis on $\{-1,+1\}^n$ is equivalent to the set of multilinear monomials, and it is an orthonormal basis of the functional space from $\{-1,+1\}^n$ to $\mathbf{R}$ with respect to the uniform distribution. Then for example, a learning algorithm for $\mathbf{AC}^0$ is designed by the following two steps [1]: First it is shown that a set of low degree monomials is enough for approximating any function in $\mathbf{AC}^0$ under the uniform distribution; Secondly, some algorithm is shown to obtain appropriate low degree polynomials (and their coefficients) from examples. We call the first step as "Fourier analysis part" and the second step as "algorithm part." Since low degree monomials are used, their learning algorithms are often referred as *Low Degree Learning Algorithms*.

In this paper, we attempt to generalize such Low Degree Learning Algorithms in the case that a distribution is far from the uniform one. We consider only the algorithm part; that is, we assume that for a given target concept class $\mathbf{C}$, some subset of Fourier basis is enough to approximate functions in $\mathbf{C}$ well. More precisely, we assume some function space $\mathcal{P}$ spanned by some subset of Fourier basis such that for every function $f \in \mathbf{C}$, we have some $g \in \mathcal{P}$ for which

$$\|f-g\|_D^2 := \mathop{\mathrm{E}}_{x\sim D}\left[(f(x)-g(x))^2\right] < \epsilon$$

for our desired ϵ. Then since the error of $\mathrm{sign}(g)$ is bounded by

$$\Pr_{x\sim D}[f(x) \neq \mathrm{sign}(g(x))] \leq \|f-g\|_D^2,$$

we can use $\mathrm{sign}(g)$ as an hypothesis to achieve the PAC-learning goal (1) *provided that g can be obtained efficiently*. In this paper, we discuss an algorithm for obtaining such g from examples. We show an algorithm that runs polynomial time (w.r.t. given parameters) if not so many instances are assinged small probabilities under D; more precisely, if $|S| < 2^{\eta n}$ for some $\eta < 1$ and for each n, where S is the set of example instances of size n whose probability under D is $< (\mathrm{poly}(n)2^n)^{-1}$.

It should be noted here that for some concept classes $\mathbf{C}$ such as $\mathbf{AC}^0$, one can define a distribution D and a function $f \in \mathbf{C}$ such that a set of (reasonably) low degree Fourier basis fails badly to approximate f under D [4, 5, 6]. Here we consider the case that such a situation does not occur.

The structure of this paper is as follows. In Section 2, we give notations and a few definitions about Fourier analysis which are used in this paper. In Section 3, we give the main results of this paper, we propose and analyze a generalization of the Low Degree Learning Algorithm.

2 Preliminaries

Let $\mathbf{N}$ be the set of all positive integers, $\mathbf{R}$ be the set of all real numbers, and $[n] := \{1,2,\ldots,n\}$. Let I be the identity matrix of any dimension.

Let $\mathcal{F}$ be the set of all functions from $\{-1,+1\}^n$ to $\mathbf{R}$. We consider $\mathcal{F}$ to be 2^n-dimensional vector space by defining $(f+g)(x) := f(x)+g(x), (\beta f)(x) :=$

$\beta f(x)$ for all $f, g \in \mathcal{F}$, $x \in \{-1,+1\}^n$ and $\beta in \mathbf{R}$. We identify a function $f \in \mathcal{F}$ to the 2^n-dimensional vector $(f(x))_{x\in\{-1,+1\}^n}$

Let D be a probability distribution over $\{-1,+1\}^n$, and U be the uniform distribution, i.e. $U(x) = 2^{-n}$ for all $x \in \{-1,+1\}^n$. We identify D to the 2^n-dimensional diagonal matrix whose entries are the probability masses of D, i.e. $D_{xx} = D(x)$ for $x \in \{-1,+1\}^n$. Note that $U \equiv 2^{-n}I$.

We define a non-canonical inner product and norm on $\mathbf{R}$.

Definition 2.1. Let D be a distribution over $\{-1,+1\}^n$, *inner product of functions $f, g \in \mathcal{F}$ with respect to D* is defined by

$$\begin{aligned}\langle f\,,\, g\rangle_D &:= \mathop{\mathrm{E}}_{x\sim D}[f(x)g(x)]\\ &= g^{\mathrm{T}}Df = f^{\mathrm{T}}Dg\end{aligned}$$

and induced (semi)norm is defined by the non-negative square root of

$$\begin{aligned}\|f\|_D^2 := \langle f\,,\, f\rangle_D &= \mathop{\mathrm{E}}_{x\sim D}\left[f(x)^2\right]\\ &= f^{\mathrm{T}}Df\end{aligned}$$

i.e. $\|f\|_D := \sqrt{\langle f\,,\, f\rangle_D}$

Definition 2.2. For a set $S \subseteq [n]$, let $\chi_S \in \mathcal{F}$ be

$$\chi_S :\ x = (x_1, x_2, \ldots, x_n) \ \mapsto\ \prod_{i=1}^{n} x_i$$

Each χ_S is *a character on the Abelian group* $\{-1,+1\}^n$ with bitwise multiplication, i.e. homomorphism to the multiplicative group of complex numbers.

Then $\{\chi_S\}_{S\subseteq[n]}$, called *Fourier basis of $\mathcal{F}$* is an orthonormal basis of $\mathcal{F}$ with inner product with respect to the uniform distribution.

Definition 2.3. For a function $f \in \mathcal{F}$, and a set $S \subseteq [n]$, *Fourier coefficient of f on S* is defined by

$$\begin{aligned}\hat{f}(S) := \langle f\,,\, \chi_S\rangle_U &= \mathop{\mathrm{E}}_{x\sim U}[f(x)\chi_S(x)]\\ &= 2^{-n}\sum_{x\in\{-1,+1\}^n} f(x)\chi_S(x)\end{aligned}$$

By orthonormality of Fourier basis, f is expanded to

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

called *Fourier expansion of f*.

3 Generalized Fourier Approximation Algorithm (GFAA)

We consider the least square approximation problem with respect to arbitary distribution. We will make an algorithm to solve the problem by examples. For a target function f and a distribution D, the algorithm tries to get a function $\tilde{g}$ which approximates f in a space spanned by some subset of Fourier basis functions. Both f and D are unknown for the algorithm, which makes queries to the example oracle $\mathrm{Ex}(f, D)$ to get information on them.

Let D be a distribution over $\{-1, +1\}^n$. Let $\mathcal{S}$ be a set of subsets of $[n]$, P be a matrix whose column vectors are a subset of Fourier basis $\{\chi_S\}_{S \in \mathcal{S}}$, $\mathcal{P}$ be the funtional subspace spanned by the column vectors of P and $K := |\mathcal{S}| = \dim \mathcal{P}$.

Formally, the problem we consider is the minimization problem

$$\min_{g \in \mathcal{P}} \|f - g\|_D^2 \tag{2}$$

Note that $\|f - g\|_D = (f - g)^{\mathrm{T}} D (f - g)$ and $g = P\beta$ for $\beta \in \mathbf{R}^K$, hence this problem can be written by

$$\min_{\beta \in \mathbf{R}^K} (f - P\beta)^{\mathrm{T}} D (f - P\beta) \tag{3}$$

Our algorithm is expected to output β of an approximate solution $\tilde{g}$ of this minimization problem.

3.1 Least square approximation under a distribution D

When $Q := P^{\mathrm{T}} D P$ is invertible, we can explicitly write the optimum solution of the problem by the standard calculation.

Lemma 3.1. Let

$$Q := P^{\mathrm{T}} D P \tag{4}$$

$$\alpha := P^{\mathrm{T}} D f \tag{5}$$

If Q is invertible, the optimum solution of (3) is written by

$$\beta^* := Q^{-1} \alpha \tag{6}$$

and the optimum solution of (2) is written by

$$g^* = P\beta^* = \sum_S \beta_S^* \chi_S \tag{7}$$

Proof. Let $C(\beta) = (f - P\beta)^{\mathrm{T}} D (f - P\beta)$, the cost function of (3),

$$\begin{aligned} C(\beta) &= f^{\mathrm{T}} D f - 2\beta P^{\mathrm{T}} D f + \beta^{\mathrm{T}} P^{\mathrm{T}} D P \beta \\ &= f^{\mathrm{T}} D f - 2\beta^{\mathrm{T}} \alpha + \beta^{\mathrm{T}} Q \beta \end{aligned}$$

The derived function of C is

$$\frac{\partial C}{\partial \beta}(\beta) = 2(Q\beta - \alpha)$$

The Hessian matrix of C is

$$\left(\frac{\partial^2 C}{\partial \beta_S \, \partial \beta_T}\right)_{S,T} = 2Q$$

and is positive definite, because Q is invertible and has decomposition $Q = (P\sqrt{D})^{\mathrm{T}}(P\sqrt{D})$ where $\sqrt{D}$ is the diagonal matrix whose entries are $\sqrt{D(x)}$ for $x \in \{-1, +1\}^n$. This means that C is strictly convex function. Hence C takes minimum value when $Q\beta = \alpha$. □

3.2 Algorithm

The key observation for making the algorithm GFAA is that each element of Q and α can be written by expectation.

Lemma 3.2. For $S, T \in \mathcal{S}$,

$$Q_{ST} = \mathop{\mathrm{E}}_{x \sim D}[\chi_S(x)\chi_T(x)] \tag{8}$$

$$\alpha_S = \mathop{\mathrm{E}}_{x \sim D}[f(x)\chi_S(x)] \tag{9}$$

Proof. For Q,

$$\begin{aligned} Q = P^{\mathrm{T}}DP &= (\chi_S)_{S\in\mathcal{S}}{}^{\mathrm{T}} D (\chi_T)_{T\in\mathcal{S}} \\ &= \left(\chi_S{}^{\mathrm{T}} D \chi_T\right)_{S,T\in\mathcal{S}} \end{aligned}$$

so an element of Q can be written by

$$Q_{ST} = \chi_S{}^{\mathrm{T}} D \chi_T = \langle \chi_S \, , \, \chi_T \rangle_D = \mathop{\mathrm{E}}_{x\sim D}[\chi_S(x)\chi_T(x)]$$

For α,

$$\alpha = P^{\mathrm{T}} D f = (\chi_S)_{S\in\mathcal{S}}{}^{\mathrm{T}} D f = \left(\chi_S{}^{\mathrm{T}} D f\right)_{S\in\mathcal{S}}$$

so an element of β can be written by

$$\alpha_S = \chi_S{}^{\mathrm{T}} D f = \langle f \, , \, \chi_S \rangle_D = \mathop{\mathrm{E}}_{x\sim D}[f(x)\chi_S(x)]$$

□

Note that Q is symmetric matrix and all diagonal entries of Q is 1, i.e. $Q_{SS} = 1$ for all $S \in \mathcal{S}$.

We can observe that each expectation can be estimated by the sample mean of the random variable. For L examples $\left(x^{(j)}, f\left(x^{(j)}\right)\right) : j \in [L]$, we estimate each element of Q and α by

$$\begin{aligned} Q_{ST} &= \mathop{\mathrm{E}}_{x\sim D}\left[\chi_S(x)\chi_T\right] \\ &\approx \frac{1}{L}\sum_{j=1}^{L} \chi_S\left(x^{(j)}\right)\chi_T\left(x^{(j)}\right) =: \tilde{Q}_{ST} \qquad (10) \\ \alpha_S &= \mathop{\mathrm{E}}_{x\sim D}\left[f(x)\chi_S(x)\right] \\ &\approx \frac{1}{L}\sum_{j=1}^{L} f\left(x^{(j)}\right)\chi_S\left(x^{(j)}\right) =: \tilde{\alpha}_S \qquad (11) \end{aligned}$$

It is sufficient for estimation of Q to estimate upper diagonal entries of Q.

Algorithm: GFAA

Input: $\epsilon, \delta > 0$, $\mathcal{S} \subseteq 2^{[n]}$, $\mathrm{Ex}(f, D)$ as the example oracle

Output: a vector $\tilde{\beta} \in \mathbf{R}^K$ of coefficients of an approximating solution $\tilde{g} = \sum_{S\in\mathcal{S}} \tilde{\beta}_S \chi_S$

1. Take L examples by $\mathrm{Ex}(f, D)$.
2. Compute $\tilde{Q}$: for $S \in \mathcal{S}, S \neq T$, compute

$$\tilde{Q}_{ST} := \frac{1}{L}\sum_{j=1}^{L} \chi_S\left(x^{(j)}\right)\chi_T\left(x^{(j)}\right)$$

   and $Q_{TS} := Q_{ST}$ and $Q_{SS} := 1$ for all $S \in \mathcal{S}$

3. Compute $\tilde{\alpha}$: for $S \in \mathcal{S}$, compute

$$\tilde{\alpha}_S := \frac{1}{L}\sum_{j=1}^{L} f\left(x^{(j)}\right)\chi_S\left(x^{(j)}\right)$$

4. Solve the linear equation $\tilde{Q}\tilde{\beta} = \tilde{\alpha}$ about $\tilde{\beta}$, and output $\tilde{\beta}$. If the equation has no solution, output $\mathbf{0} \in \mathbf{R^K}$.

We leave to determine the number of examples L, which is preferable to be within a polynomial of n, ϵ, δ and K.

By Lemma 3.1, this algorithm outputs the optimum solution of (6) in an ideal case that Q is invertible and no numerical error occurs. In this section, we do not treat the case that Q is not invertible.

We ensure that the algorithm works when Q has no very small eigenvalue.

Theorem 3.3. Let μ be a positive value, which may depend on n, D, K and P, if $Q = P^{\mathrm{T}}DP$ has no eigenvalue less than μ, then the number of examples

$$L \leq O\left(K^6\epsilon^{-1}\mu^{-4}\ln\left(K\delta^{-1}\right)\right) \tag{12}$$

is sufficient for GFAA to output coefficients of $\tilde{g} \in \mathcal{P}$ such that

$$\|f - \tilde{g}\|_D^2 \leq \|f - g^*\|_D^2 + \epsilon \tag{13}$$

with probability not less than $1-\delta$, and the algorithm works in polynomial time of n, K, ϵ^{-1}, $\delta{-}1$ and μ^{-1}.

Particularly,

$$O\left(K^3\epsilon^{-1}\mu^{-2}\ln\left(K\delta^{-1}\right)\right) \tag{14}$$

labeled examples are sufficient, i.e. other examples are not needed to be labeled.

Before going on to the proof, we will give the summary of it.

By Lemma 3.1, the optimum solution is $g^* = PQ^{-1}\alpha$. If $\tilde{Q}$ is also invertible, the algorithm outputs coefficients of $\tilde{g} = P\tilde{Q}^{-1}\tilde{\alpha}$, otherwise it outputs $\mathbf{0}$. Note that Q is invertible because Q has no zero eigenvalue.

Let $E_Q := \tilde{Q} - Q$ and $e_\alpha := \tilde{\alpha} - \alpha$. Note that E_Q is symmetric.

We will bound the difference between the minimum distance $\min_{g\in\mathcal{P}} \|f - g\|_D$ and $\|f - \tilde{g}\|_D$, where $\tilde{g}$ is an approximate solution produced by GFAA, by E_Q and e_α. The later calculation gives the upper bound:

$$\|f - \tilde{g}\|_D^2 - \|f - g^*\|_D^2 \leq 3K\left(\frac{2\sqrt{K}}{\mu^2}\|E_Q\| + \frac{2}{\mu}\|e_\alpha\|\right)^2 \tag{15}$$

assuming that $\|E_Q\|, \|e_\alpha\| \leq \mu/2$.

The algorithm achieves the accuracy (13) when the following two conditions are satisfied.

$$\|E_Q\|^2 \leq \frac{\mu^4}{48K^2}\,\epsilon \quad \text{and} \tag{16}$$

$$\|e_\alpha\|^2 \leq \frac{\mu^2}{48K}\,\epsilon \tag{17}$$

We determine the number of examples L such that the conditions are satisfied with probability grater than $1 - \delta$.

To obtain (15), we make a few assumptions for simplicity. They are justified when GFAA takes a sufficient number of examples to satisfy (16), (17).

Claim 1. Assume that $\|\tilde{g} - g^*\|_D \leq 1$, then

$$\|f - \tilde{g}\|_D^2 - \|f - g^*\|_D^2 \leq 3\|\tilde{g} - g^*\| \tag{18}$$

Proof.

$$\begin{aligned}\|f - \tilde{g}\|_D^2 - \|f - g^*\|_D^2 &= -2\left\langle f\,,\,\tilde{g}\right\rangle_D^2 + \|\tilde{g}\|_D^2 + 2\left\langle f\,,\,g^*\right\rangle_D - \|g^*\|_D^2 \\ &= 2\left\langle f\,,\,g^* - \tilde{g}\right\rangle_D + \left\langle \tilde{g} + g^*\,,\,\tilde{g} - g^*\right\rangle_D \\ &= \left\langle 2f - \tilde{g} - g^*\,,\,g^* - \tilde{g}\right\rangle_D\end{aligned}$$

By Cauchy–Schwarz inequality, and triangle inequality,

$$\begin{aligned}\langle 2f - \tilde{g} - g^* \, , \, g^* - \tilde{g}\rangle_D &\le \|2f - \tilde{g} - g^*\|_D \, \|g^* - \tilde{g}\|_D \\ &\le (2\,\|f - g^*\|_D + \|\tilde{g} - g^*\|_D)\, \|\tilde{g} - g^*\|_D\end{aligned}$$

and

$$(2\,\|f - g^*\|_D + \|\tilde{g} - g^*\|_D)\, \|\tilde{g} - g^*\|_D \le \left(2 + \|\tilde{g} - g^*\|_D^2\right) \|\tilde{g} - g^*\|_D^2$$

because

$$\|f - g^*\|_D^2 \le \|f - \mathbf{0}\|_D^2 = \|f\|_D^2 = 1$$

By assumption $\|\tilde{g} - g^*\|_D^2 \le 1$, we obtain (18). □

Claim 2.

$$\|\tilde{g} - g^*\|_D^2 \le K \left\|\tilde{Q}^{-1}\tilde{\alpha} - Q^{-1}\alpha\right\|^2 \tag{19}$$

Proof.

$$\begin{aligned}\|\tilde{g} - g^*\|_D^2 &= (\tilde{g} - g^*)^{\mathrm{T}} D\, (\tilde{g} - g^*) \\ &= \left(\tilde{Q}^{-1}\tilde{\alpha} - Q^{-1}\alpha\right)^{\mathrm{T}} P^{\mathrm{T}} D P \left(\tilde{Q}^{-1}\tilde{\alpha} - Q^{-1}\alpha\right) \\ &= \left(\tilde{Q}^{-1}\tilde{\alpha} - Q^{-1}\alpha\right)^{\mathrm{T}} Q \left(\tilde{Q}^{-1}\tilde{\alpha} - Q^{-1}\alpha\right) \\ &\le \|Q\| \left\|\tilde{Q}^{-1}\tilde{\alpha} - Q^{-1}\alpha\right\|^2\end{aligned}$$

Since Q is a positive definite symmetric matrix, and all diagonal entries of Q are 1, $\|Q\| \le \mathrm{tr}(Q) = K$. Thus, we obtain (19). □

Claim 3.

$$\left\|\tilde{Q}^{-1}\tilde{\alpha} - Q^{-1}\alpha\right\| \le \sqrt{K} \left\|\tilde{Q}^{-1} - Q^{-1}\right\| + \left\|\tilde{Q}^{-1}\right\| \|e_\alpha\| \tag{20}$$

Proof. By $\tilde{\alpha} = \alpha + e_\alpha$, we have

$$\begin{aligned}\left\|\tilde{Q}^{-1}\tilde{\alpha} - Q^{-1}\alpha\right\| &= \left\|\left(\tilde{Q}^{-1} - Q^{-1}\right)\alpha + \tilde{Q}^{-1}e_\alpha\right\| \\ &\le \left\|\tilde{Q}^{-1} - Q^{-1}\right\| \|\alpha\| + \left\|\tilde{Q}^{-1}\right\| \|e_\alpha\|\end{aligned}$$

Note that $|\alpha_S| \le 1$ for all $S \in \mathcal{S}$ because α_S is the expectation of ±1-valued random variable. Hence

$$\|\alpha\|^2 = \sum_{t=1}^{K} \alpha_{S_t}^2 \le K$$

Thus, we obtain (20). □

We bound each norm of (20). For simplicity, we assume that $\|E_Q\|, \|e_\alpha\| \le \mu/2$, which is also justified by (16), (17).

Claim 4. Assume that Q is invertible and $\|E_Q\| \leq \mu/2$, then

$$\left\|\tilde{Q}^{-1}\right\| \leq \frac{2}{\mu} \tag{21}$$

Proof. Let λ be the smallest eigenvalue of Q, and $lam\tilde{b}da$ be that of $\tilde{Q}$. Since Q, $\tilde{Q}$ and E_Q are symmetric,

$$\begin{aligned} \min_z z^{\mathrm{T}} Q z &= \|Q^{-1}\| = \frac{1}{\lambda} \\ \min_z z^{\mathrm{T}} Q^{-1} z &= \left\|\tilde{Q}^{-1}\right\| = \frac{1}{\tilde{\lambda}} \\ \min_z z^{\mathrm{T}} E_Q z &= \|E_Q\| \end{aligned}$$

where z moves on all over the unit sphere $\left\{z \in \mathbf{R}^K : \|z\| = 1\right\}$.

Thus,

$$\begin{aligned} \tilde{\lambda} &= \min_z \left(z^{\mathrm{T}} Q z + z^{\mathrm{T}} E_Q z\right) \\ &\geq \min_z z^{\mathrm{T}} Q z - \max_z z^{\mathrm{T}} E_Q z \\ &= \lambda - \|E_Q\| \end{aligned}$$

By the conditions $\lambda \geq \mu$ and $\|E_Q\| \leq \mu/2$, we have

$$\left\|\tilde{Q}^{-1}\right\| = \frac{1}{\lambda} \leq \frac{1}{\lambda - \|E_Q\|} \leq \frac{1}{\mu - \mu/2} \leq \frac{2}{\mu}$$

□

Claim 5.

$$\left\|\tilde{Q}^{-1} - Q^{-1}\right\| \leq \frac{2}{\mu^2} \|E_Q\| \tag{22}$$

Proof. We use expansion to matrix power series. By the assumption,

$$\begin{aligned} \tilde{Q}^{-1} &= (Q + E_Q)^{-1} = Q^{-1}\left(I + Q^{-1} E_Q\right)^{-1} \\ \|Q^{-1} E_Q\| &\leq \|Q^{-1}\| \, \|E_Q\| \leq \frac{1}{\mu}\frac{\mu}{2} = \frac{1}{2} < 1 \end{aligned}$$

Hence $\tilde{Q}^{-1}$ is expanded by

$$\tilde{Q}^{-1} = Q^{-1} \sum_{k=0}^{\infty} \left(-Q^{-1} E_Q\right)^k Q^{-1} + Q^{-1} \sum_{k=1}^{\infty} \left(-Q^{-1} E_Q\right)^k$$

Therefore,

$$\begin{aligned} \left\|\tilde{Q}^{-1} - Q^{-1}\right\| &= \left\| Q^{-1} \sum_{k=1}^{\infty} \left(-Q^{-1} E_Q\right)^k \right\| \leq \|Q^{-1}\| \sum_{k=1}^{\infty} \left\|Q^{-1} E_Q\right\|^t \\ &= \frac{\|Q^{-1}\| \, \|Q^{-1} E_Q\|}{1 - \|Q^{-1} E_Q\|} \leq \frac{\|Q^{-1}\|^2}{1 - \|Q^{-1} E_Q\|} \|E_Q\| \\ &\leq \frac{2}{\mu^2} \|E_Q\| \end{aligned}$$

□

By (21) and (22), (20) is bounded by

$$\left\|\tilde{Q}^{-1}\tilde{\alpha} - Q^{-1}\alpha\right\| \leq \frac{2\sqrt{K}}{\mu^2}\left\|E_Q\right\| + \frac{2}{\mu}\left\|e_\alpha\right\|$$

Hence (20) is bounded by

$$\|\tilde{g} - g^*\|_D^2 \leq K\left(\frac{2\sqrt{K}}{\mu^2}\|E_Q\| + \frac{2}{\mu}\|e_\alpha\|\right)^2$$

If (16) and (17) are satisfied, the assumption of Claim 1 is also satisfied since $\|\tilde{g} - g^*\|_D^2 \leq \epsilon/3 \leq 1$. Hence we obtain (17) by evaluation of (18).

We bound the number of examples L which is sufficient for the algorithm to satisfy both (16) and (17) with probability not less than than $1 - \delta$.

We use a well-known fact about relationship of matrix norms. (See [8]) For a matrix $A := [a_{ij}]$, the following inequality holds:

$$\|A\|^2 \leq \sum_{i,j} a_{ij}^2 \tag{23}$$

Claim 6. For Gereralized Fourier Approximation Algorithm,

$$L \leq O\left(K^4\left(\epsilon_Q^{-1} + \epsilon_\alpha^{-1}\right)\ln\left(K\delta^{-1}\right)\right) \tag{24}$$

examples is sufficient to satisfy the following two conditions:

$$\|E_Q\|^2 \leq \epsilon_Q \qquad \|e_\alpha\|^2 \leq \epsilon_\alpha \tag{25}$$

with probability not less than $1 - \delta$.

Particularly,

$$O\left(K^3\epsilon^{-1}\mu^{-2}\ln\left(K\delta^{-1}\right)\right) \tag{26}$$

labeled examples are sufficient.

Proof. For convenience, we denote each element of maxtrices and vectors by subscripted numbers, e.g. Q_{st} where $s, t \in [K]$.

By (23), it is sufficient to bound the probability such that one of the events

$$\|E_Q\|^2 \leq \sum_{s,t=1}^{K}\left(\tilde{Q}_{st} - Q_{st}\right)^2 \quad \text{and} \tag{27}$$

$$\|e_\alpha\|^2 = \sum_{t=1}^{K}\left(\tilde{\alpha}_t - \alpha_t\right)^2 \tag{28}$$

does not occur.

Let $\epsilon'_Q, \epsilon'_\alpha, \delta'_Q, \delta'_\alpha$ be positive values. If both

$$\Pr\left[\left|Q_{st} - \tilde{Q}_{st}\right| > \epsilon'_Q\right] \leq \delta'_Q \quad \text{and} \tag{29}$$

$$\Pr\left[\left|\alpha_t - \tilde{\alpha}_t\right| > \epsilon'_\alpha\right] \leq \delta'_\alpha \tag{30}$$

hold, then both

$$\Pr\left[\exists s, \exists t \in [K] \text{ s.t. } \left|Q_{st} - \tilde{Q}_{st}\right| > \epsilon'_Q\right] \leq K^2 \delta'_Q \quad \text{and}$$
$$\Pr\left[\exists \in [K] \text{ s.t. } \left|\alpha_t - \tilde{\alpha}_t\right| > \epsilon'_\alpha\right] \leq K\delta'_\alpha$$

also hold. For simplicity, we use a relaxed upper bound about Q. Hence both

$$\Pr\left[\sum_{s,t=1}^{K} \left(Q_{st} - \tilde{Q}_{st}\right)^2 > K^2 \left(\epsilon'_Q\right)^2\right] \leq K^2 \delta'_Q \quad \text{and}$$
$$\Pr\left[\sum_{t=1}^{K} \left(\alpha_t - \tilde{\alpha}_t\right)^2 > K \left(\epsilon'_\alpha\right)^2\right] \leq K\delta'_\alpha$$

hold.

Let

$$\epsilon'_Q := \sqrt{\epsilon_Q}/K, \qquad \delta'_Q := \delta/2K^2,$$
$$\epsilon'_\alpha := \sqrt{\epsilon_\alpha/K}, \qquad \delta'_\alpha := \delta/2K$$

Then both

$$\Pr\left[\|E_Q\|^2 > \epsilon_Q\right] \leq \Pr\left[\sum_{s,t=1}^{K} \left(Q_{st} - \tilde{Q}_{st}\right)^2 > \epsilon_Q\right] \leq \frac{\delta_Q}{2} \quad \text{and}$$
$$\Pr\left[\|E_\alpha\|^2 > \epsilon_\alpha\right] = \Pr\left[\sum_{t=1}^{K} \left(\alpha_t - \tilde{\alpha}_t\right)^2 > \epsilon_\alpha\right] \leq \frac{\delta_\alpha}{2}$$

hold. By union bound, the probability such that one of (25) is not satisfied is less than δ. Thus, (25) is satisfied with probability not less than than $1 - \delta$.

We have left to bound (29) and (30). For each $s, t \in [K]$, let

$$a_{st}^{(j)} := \chi_{S_s}\left(x^{(j)}\right)\chi_{S_t}\left(x^{(j)}\right) \quad \text{and}$$
$$b_t^{(j)} := f\left(x^{(j)}\right)\chi_{S_t}\left(x^{(j)}\right)$$

Then, Q_{st} and α_t are written by $Q_{st} = \mathrm{E}\left[a_{st}^{(j)}\right]$, $\alpha_t = \mathrm{E}\left[b_t^{(j)}\right]$.

$a_{st}^{(j)}$, $b_t^{(j)}$ are $pm1$ valued independent random values for each other j. Hence we can bound (29) and (30) by Höffding's ineqality as

$$\Pr\left[\left|Q_{st} - \tilde{Q}_{st}\right| > \epsilon'_Q\right] \leq 2\exp\left(-\frac{2L(\epsilon'_Q)^2}{4}\right) = 2\exp\left(-\frac{L\epsilon_Q}{2K^4}\right) \quad \text{and}$$
$$\Pr\left[\left|\alpha_t - \tilde{\alpha}_t\right| > \epsilon'_\alpha\right] \leq 2\exp\left(-\frac{2L(\epsilon'_\alpha)^2}{4}\right) = 2\exp\left(-\frac{L\epsilon_\alpha}{2K^2}\right)$$

Therefore, the number of examples

$$L \leq O\left(\frac{K^4}{\epsilon_Q}\ln\left(\frac{K}{\delta}\right) + \frac{K^2}{\epsilon_\alpha}\ln\left(\frac{K}{\delta}\right)\right)$$

are sufficient to satisfy both (24). □

Set

$$\begin{aligned}\epsilon_Q &:= \frac{\mu^4}{48K^2}\,\epsilon \quad \text{and} \\ \epsilon_\alpha &:= \frac{\mu^2}{48K}\,\epsilon\end{aligned}$$

Then

$$O\left(K^6\epsilon^{-1}\mu^{-4}\ln\left(K\delta^{-1}\right)\right)$$

examples are sufficient to estimate Q with required accuracy, and they are not needed to be labeled. And

$$O\left(K^3\epsilon^{-1}\mu^{-2}\ln\left(K\delta^{-1}\right)\right)$$

labeled examples are sufficient to estimate α.

Finally, we observe that the algorithm works in polynomial time with respect to n, K and L. Since $\tilde{Q}$ has K^2 elements and $\tilde{\alpha}$ has K elements and the algorithm queries L examples, the number of elementary operations to calculate all elements of $\tilde{Q}$ and $\tilde{\alpha}$ is in polynomial of K and L. And the numbers of elementary operations for product and inverse operation of K dimensional vectors and matrices is also in polynomial of K. Elementary operations can be done in polynomial time with respect to the length of variables.

Since L is within polynomial of K, ϵ^{-1}, δ^{-1} and μ^{-1}, the algorithm works in polynomial time of these variables.

3.3 Relating to Distribution

If distribution D has no small probability mass, we ensure that Q has no small eigenvalue, and the algorithm works.

Theorem 3.4. Let $\mu_{\min} := 2^n \min_x D(x) > 0$, then the number of examples

$$L \leq O\left(K^6\epsilon^{-1}\mu_{\min}^{-4}\ln\left(K\delta^{-1}\right)\right) \tag{31}$$

is sufficient for GFAA to output coefficients of $\tilde{g} \in \mathcal{P}$ such that

$$\|f - \tilde{g}\|_D^2 \leq \|f - g^*\|_D^2 + \epsilon \tag{32}$$

with probability not less than than $1-\delta$, and the algorithm works in polynomial time of n, K, ϵ^{-1}, δ^{-1} and μ^{-1}.

It is sufficient to prove that Q has no small eigenvalue.

Lemma 3.5. Let $\lambda_{\min}$ be the smallest eigenvalue of Q, then $\lambda_{\min} \geq \mu_{\min}$.

Proof. By the condition, $D(x)$ can be written by $D(x) = D'(x) + \mu_{\min}$, where $D'(x) \geq 0$ for all $x \in \{-1, +1\}^n$. We identify D' as diagonal matrix as well as D. Then Q can be written by

$$\begin{aligned} Q &= P^{\mathrm{T}} D P = P^{\mathrm{T}} D' P + \mu P^{\mathrm{T}} P \\ &= P^{\mathrm{T}} D' P + 2^n \mu_{\min} I \end{aligned}$$

$2^{-n} P^{\mathrm{T}} P = I$ follows by orthonormality of Fourier basis. Let $Q' = P^{\mathrm{T}} D' P$, Q' is positive semi-definite symmetric matrix because of $D'(x) \geq 0$. Hence

$$\begin{aligned} \lambda_{\min} &= \min_z z^{\mathrm{T}} Q z = \min_z \left(z^{\mathrm{T}} Q' z + \mu_{\min} z^{\mathrm{T}} z \right) \\ &\geq \min_z z^{\mathrm{T}} Q' z + \mu_{\min} \geq \mu_{\min} \end{aligned}$$

where $\|z\| = 1$. □

We can relax the condition that the distribution has 'no' very small probability mass to the condition that the distribution has 'few' very small probability mass.

Theorem 3.6. Assume that $K \leq 2^{o(n)}$. Let $\mu > 0$ and $S := \{x : 2^n D(x) \leq \mu\}$. If there exists $0 < \eta < 1$ such that $|S| \leq 2^{\eta n}$, then the number of examples

$$L \leq O\left(K^6 \epsilon^{-1} \mu^{-4} \ln \left(K \delta^{-1} \right) \right) \tag{33}$$

is sufficient for GFAA to output coefficients of $\tilde{g} \in \mathcal{P}$ such that

$$\|f - \tilde{g}\|_D^2 \leq \|f - g^*\|_D^2 + \epsilon \tag{34}$$

with probability not less than than $1 - \delta$, and the algorithm works in polynomial time of n, K, ϵ^{-1}, δ^{-1} and μ^{-1}.

The proof is similar to that of Lemma 3.5.

Lemma 3.7. $\lambda_{\min} \geq \mu \left(1 - K 2^{-(1-\eta)n} \right)$.

Proof. Let $\bar{S} := \{-1, +1\}^n \backslash S$, the complement set of S and $\bar{U}$ be the uniform distribution over $\bar{S}$. By the condition, D can be written by $D(x) = D'(x) + 2^{-n} \mu |\bar{S}| \bar{U}_S(x)$ and $D'(x) \geq 0$ for all $x \in \{-1, +1\}^n$. We identify D' and $\bar{U}$ as diagonal matrix as well as D. Then Q can be written by

$$Q = P^{\mathrm{T}} D P = P^{\mathrm{T}} D' P + 2^{-n} \mu |\bar{S}| P^{\mathrm{T}} \bar{U} P$$

Let $Q' = P^{\mathrm{T}} D' P$ and $\bar{Q} = P^{\mathrm{T}} \bar{U} P$. Note that Q' and $\bar{Q}$ are positive semi-definite symmetric matrices because of $D'(x), \bar{U} \geq 0$. Hence $\bar{\lambda}_{\min} := \min_z z^{\mathrm{T}} \bar{Q} z$ is the smallest eigenvalue of $\bar{Q}$, and

$$\begin{aligned} \lambda_{\min} &= \min_z z^{\mathrm{T}} Q z = \min_z \left(z^{\mathrm{T}} Q' z + 2^{-n} \mu |\bar{S}| z^{\mathrm{T}} \bar{Q} z \right) \\ &\geq \min_z z^{\mathrm{T}} Q' z + 2^{-n} \mu |\bar{S}| \bar{\lambda}_{\min} \\ &\geq 2^{-n} \mu |\bar{S}| \bar{\lambda}_{\min} \end{aligned}$$

where $\|z\| = 1$.

We remain to show that $\bar{\lambda}_{\min}$ is not too small. The following inequality follows by the next lemma.

$$\lambda_{\min} \geq 2^{-n}\mu|\bar{S}|\bar{\lambda}_{\min} \geq \mu\left(1 - K2^{-(1-\eta)n}\right) \tag{35}$$

□

Lemma 3.8. $\bar{\lambda}_{\min}|\bar{S}| \geq 2^n\left(1 - K2^{-(1-\eta)n}\right)$

Proof. By Lemma 3.2, $\bar{Q}_{TT'} = \mathrm{E}_{x\in\bar{U}}\left[\chi_T(x)\chi_{T'}(x)\right]$.

We show that $\bar{Q}$ is close to the identity matrix, i.e. all off-diagonal entries of $\bar{Q}$ are small.

Since Fourier basis is closed under multiplication, without loss of generality, off-diagonal entries of $\bar{Q}$ can be written by $\mathrm{E}_{x\in\bar{U}}\left[\chi_R(x)\right]$ by some $R \neq \emptyset$. Since χ_R is orthogonal to $\chi_\emptyset = \mathbf{1}$, there are same numbers of $x \in \{-1,+1\}^n$ such that $\chi_R(x) = +1$ and $\chi_R(x) = -1$. For any S of fixed size s and any $R \neq \emptyset$, let S' be a set of same size s such that for all $x \in S'$ $\chi_R(x) = -1$, then

$$|\bar{S}| \mathop{\mathrm{E}}_{x\in\bar{U}}\left[\chi_R(x)\right] = \sum_{x\in\bar{S}}\chi_R(x) \leq \sum_{x\in\bar{S}'}\chi_R(x)$$

If $\chi_R(x_+) = +1$ for some $x_+ \in S$, there must be $x_- \in \bar{S}$ such that $\chi_R(x_-) = -1$. Hence $\sum_{x\in\bar{S}}\chi_R(x)$ can not exceed $\sum_{x\in\bar{S}'}\chi_R(x)$. Thus,

$$\mathop{\mathrm{E}}_{x\in\bar{U}}\left[\chi_R(x)\right] \leq \frac{1}{|\bar{S}|}\sum_{x\in\bar{S}'}\chi_R(x) = \frac{1}{|\bar{S}|}\cdot(+1)\cdot|S| = \frac{|S|}{|\bar{S}|}$$

By similar way, we can show

$$-\mathop{\mathrm{E}}_{x\in\bar{U}}\left[\chi_R(x)\right] \leq \frac{|S|}{|\bar{S}|}$$

Therefore, we obtain

$$|\bar{Q}_{TT'}| \leq \left|\mathop{\mathrm{E}}_{x\in\bar{U}}\left[\chi_R(x)\right]\right| \leq \frac{|S|}{|\bar{S}|}$$

By Geršgorin disc theorem (See [8]),

$$\bar{\lambda}_{\min} \geq \min_T\left(|\bar{Q}_{TT}| - \sum_{T'\neq T}|\bar{Q}_{TT'}|\right) \geq 1 - (K-1)\frac{|S|}{|\bar{S}|}$$

Hence,

$$\begin{aligned}\bar{\lambda}_{\min}|\bar{S}| &\geq |\bar{S}| - (K-1)|S| = 2^n - K|S| \\ &\geq 2^n\left(1 - K2^{-(1-\eta)n}\right)\end{aligned}$$

□

4 Conclusions

We made the algorithm GFAA which is an extension of the Low Degree Algorithm and showed that if the smallest eigenvalue of the matrix whose values are Fourier coefficients of the data distribution is not too small. We also showed that the matrix has no very small eigenvalue if the distribution has few very small probability mass. But the algorithm GFAA may not work when this condition is not satisfied. Our furthur motivation is to modify GFAA to work under arbitary distributions. We consider whether regularization techniques work.

Acknowledgment

I thank Prof. Osamu Watanabe for helpful discussions and several valuable comments.

References

[1] N. Linial, Y. Mansour and N. Nisan. Constant depth circuits, Fourier transform and learnability In *Proc. 30th FOCS*, pages 574–579, 1989

[2] Y. Mansour, An $O(n^{\log \log n})$ learning algorithm for DNF under the uniform distribution, In *Proc. COLT*, pages 53–61, 1992

[3] J. Jackson, An Efficient Membership-Query Algorithm for Learning DNF with Respect to the Uniform Distribution, In *Proc. 35th FOCS*, pages 42–53, 1994

[4] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry*, MIT Press, 1968

[5] R. O'Donnel and R. Servedio. New degree bounds for polynomial threshold functions. In *Proc. 35th STOC*, pages 325–334, 2003

[6] A. Sherstov. Separating $\mathbf{AC}_0$ from depth-2 majority circuits. In *Proc. 39th STOC*, pages 294–301, 2007

[7] L. Valiant. A theory of learnable. In *Communications of ACM 27*, pages 1134–1142, 1984

[8] R. Horn and C. Johnson. *Matrix Analysis.* Cambridge University Press, 1990