6.3.2008 **PAC Model** [Valiant '84]

We consider only the Probably Approximately Correct (PAC) model under the uniform distribution A learning problem is given by a <u>concept class</u> C of functions $f:\{0,1\}^n \to \{-1,1\}$.

A learning algorithm is (typically) randomized algorithm with:

- Input: accuracy parameter $\varepsilon > 0$.
- Either: ① Random examples $(x, f(x))$ for x chosen uniformly, or
  ② Membership queries to f. (where $f \in C$)
- Output: a function $h:\{0,1\}^n \to \{-1,1\}$, known as the hypothesis, described by a circuit.
- Goal: h should be ε-close to f with prob. 99%.

<u>Remarks</u>:
- General-PAC learning requries the algorithm to work under any dist. This affects both choice of random samples and the closeness of h and f. This typically very hard.
- Often one requires h to be of the same "form" as C. This is called proper learning and is also typically much harder.
- One can always check if the hypothesis h is ε-close to f: simply compare them on enough random examples. This also means that one can always boost the 99% correctness to $1-\delta$ by paying an extra $O(\log 1/\delta)$ factor.

Learning Using Spectral Concentration

<u>Def</u>: For a family $\mathcal{S}$ of subsets of $[n]$, we say that f is <u>ε-concentrated</u> on $\mathcal{S}$ if $\sum_{S \notin \mathcal{S}} \hat{f}(S)^2 \le \varepsilon$.

<u>Claim</u>: If $f:\{0,1\}^n \to \{-1,1\}$ is ε-concentrated on $\mathcal{S}$ then $g:\{0,1\}^n \to \mathbb{R}$ given by $g = \sum_{S \in \mathcal{S}} \hat{f}(S)\chi_S$ satisfies $\|f-g\|_2 \le \varepsilon$.

<u>Proof</u>: $f-g = \sum_{S \notin \mathcal{S}} \hat{f}(S)\chi_S$ and hence, $\|f-g\|_2^2 = \sum_{S \notin \mathcal{S}} \hat{f}(S)^2 \le \varepsilon$. ∎

<u>Claim</u>: Let $f:\{0,1\}^n \to \{-1,1\}$ and $g:\{0,1\}^n \to \mathbb{R}$ satisfy $\|f-g\|_2^2 \le \varepsilon$. Then, $h:\{0,1\}^n \to \{-1,1\}$ given by $h(x) = \mathrm{sign}(g(x))$ is ε-close to f.

<u>Proof</u>: For each x such that $f(x) \ne h(x)$ we must have $(f(x)-g(x))^2 \ge 1$. Now use $\|f-g\|_2^2 = \mathbb{E}_x\left[(f(x)-g(x))^2\right]$. ∎

Thm: [Linial, Mansour, Nisan, 89'] Suppose we know a set S on which f is $\varepsilon/2$-concentrated. Then, in time $\mathrm{poly}(|S|, 1/\varepsilon, n)$ using random examples, we can output a function h that is ε-close to f w.p. $\geq 99\%$.

Proof: The algorithm – for each $s \in S$ estimate $\hat{f}(s)$ to within $\pm\sqrt{\frac{\varepsilon}{16|S|}}$ and let $\tilde{f}_s$ be the estimate. Let $\tilde{g} = \sum_{s\in S} \tilde{f}_s \cdot \chi_s$. Now output $h = \mathrm{sign}(\tilde{g})$.

This takes $\mathrm{poly}(|S|, n, 1/\varepsilon)$ time. By the previous claim, it suffices to show that $\|f-\tilde{g}\|_2^2 \leq \varepsilon$. Define $g = \sum_{s\in S} \hat{f}(s)\chi_s$. Then, $\|f-\tilde{g}\|_2 \underset{\text{triangle inequality}}{\leq} \|f-g\|_2 + \|g-\tilde{g}\|_2 \leq$

$\leq \sqrt{\varepsilon/2} + \sqrt{\sum_{s\in S}(\hat{f}(s)-\tilde{f}_s)^2} \leq \sqrt{\varepsilon/2} + \sqrt{|S|\cdot\frac{\varepsilon}{16|S|}} \leq \sqrt{\varepsilon}$. ∎

Cor: (The low degree algorithm [LMN]) If all $f \in C$ is $\varepsilon/2$-concentrated on $S = \{s : |s| \leq d\}$ then C can be learned in time $n^{O(d)} \cdot \mathrm{poly}(1/\varepsilon)$ using random examples.

Cor [Kushilevitz Mansour '91]: If any $f \in C$ $\varepsilon/4$-concentrated on some set of size $\leq M$ then C can be learned in time $\mathrm{poly}(M, 1/\varepsilon, n)$ using membership queries.

Proof: Assume we are given some $f \in C$ and let S be a set of size $\leq M$ on which it is $\varepsilon/4$-concentrated. Let $S' = \{s \in S \mid \hat{f}(s)^2 \geq \frac{\varepsilon}{4M}\}$. Then f is $\varepsilon/2$-concentrated on S'. Using the G-L algorithm we can find a set L containing all s s.t. $\hat{f}(s)^2 \geq \frac{\varepsilon}{4M}$ in time $\mathrm{poly}(M, 1/\varepsilon, n)$. In particular, $S' \subseteq L$ so f is $\frac{\varepsilon}{2}$-concentrated on L. Now apply [LMN] to find h that is ε-close to f. ∎

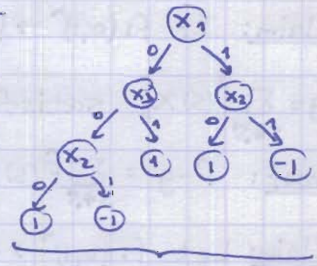
Thm: The class $C = \{f : I(f) \leq d\}$ is learnable in time $n^{O(d/\varepsilon)}$ using random examples.

Proof: Recall that $I(f) = \sum_s |s| \cdot \hat{f}(s)^2$. Notice that $I(f) \leq d$ implies that $\sum_{|s| > d/\varepsilon} \hat{f}(s)^2 \leq \varepsilon$. Hence, we can use the Low Degree Alg. with $S = \{s : |s| \leq d/\varepsilon\}$.

Learning Desision Trees

Remarks: We assume w.l.o.g. that on any path each variable appears at most once.

depth = 3



size = #leaves = 5.

Notice that the decision trees are universal, i.e., any Boolean function can be written as a decision tree (of depth $\leq n$ and size $\leq 2^n$).

A decision tree of depth d has size $\leq 2^d$.

Prop: Assume $f:\{0,1\}^n \to \{-1,1\}$ is computable by a depth-d decision tree. Then,

1. f is of degree $\le d$ (i.e., $\sum_{|S|>d} \hat{f}(S)^2 = 0$).

2. All Fourier coeff. of f are integer multiples of 2^{-d}.

3. The number of nonzero coeff. is at most 4^d.

Proof: For each path P in the decision tree, let 1_P be the indicator function of P. Then, we can write $f = \sum_P \underbrace{f(P)}_{\pm 1} \cdot 1_P$ because the paths define a partition of $\{0,1\}^n$. Because 1_P is the AND of at most d literals, its Fourier coeff. are multiples of 2^{-d}, and is of degree at most d. (since, e.g., $\mathrm{AND}(x_1, \overline{x_3}, x_2) = \left(\frac{1-x_{\{1\}}}{2}\right)\left(\frac{1+x_{\{3\}}}{2}\right)\left(\frac{1-x_{\{2\}}}{2}\right)$).

This proves 1&2. 3 follows from 2. ∎

Cor: 1. Depth-d DTs are exactly learnable (i.e., with $\varepsilon=0$) in time $n^{O(d)}$ using random examples.

2. Depth-d DTs are exactly learnable in time $\mathrm{poly}(n, 2^d)$ using membership queries.

3. Depth-$O(\log n)$ DTs are exactly learnable in time $\mathrm{poly}(n)$ using membership queries.

Proof: Estimate all nonzero Fourier coeff. to within $\pm 2^{-d}/4$ and round to nearest multiple of 2^{-d}. ∎

Observation: Any DT of size L is ε-close to a DT of depth $\log(L/\varepsilon)$.

Proof: Cut everything deeper than $\log(L/\varepsilon)$. The resulting DT differs from the original on a random input x w.p. $\le L \cdot 2^{-\log(L/\varepsilon)} = \varepsilon$. ∎

Cor: DTs of size L are 4ε-concentrated on a set of size $4^{\log(L/\varepsilon)} = (L/\varepsilon)^2$ of Fourier coeff of level $\le \log(L/\varepsilon)$.

Hence, DTs of size L can be learned in time $\mathrm{poly}(L, n, 1/\varepsilon)$ using queries and time $n^{O(\log(L/\varepsilon))}$ using random examples.

Remark: These are the best known results. It is an open question to do poly-size DT in time $\mathrm{poly}(n)$ using random examples.

Learning DNFs

Def: A DNF is a disjunction $\phi = T_1 \vee T_2 \vee \dots \vee T_L$ where each term is a conjunction of literals (e.g., $\underbrace{(x_1 \wedge \overline{x_5}) \vee (x_3 \wedge x_4 \wedge x_5)}_{\text{size}=2,\ \text{width}=3}$). The size of a DNF is the number of terms L, and the width of a DNF is the maximal # of literals in a term.

Remark: Any DT of size L can be written as a DNF of size L. In other words, DNF-size(f) $\le$ DT-size(f). Similarly, DNF-width(f) $\le$ DT-depth(f).

As we saw in homework, a width w DNF f has $I(f) \le 2w$ and hence is ε-concentrated on levels $\le \frac{2w}{\varepsilon}$. Our goal is to improve this to $O(w \cdot \log 1/\varepsilon)$.

Thm: If f is computable by a width-w DNF then for any $d \ge 5$,

$$\sum_{|S| \ge 20dw} \hat{f}(S)^2 \le 2^{-d+1}.$$

Def: A random p-restriction is a pair (I, x) where $I \subseteq [n]$ is chosen by including each coordinate w.p. p and $x \in \{0,1\}^{\bar{I}}$ is chosen uniformly.

Thm [Hastad's switching lemma '86]: If f is computable by a width-w DNF and (I,x) is a random p-restriction, then $\Pr[\text{DT-depth}(f_{x \to \bar{I}}) > d] \le (5pw)^d$.

Example: $p = \frac{1}{10w}$, then we get that w.p. $\ge 3/4$ the restriction has DT of depth ≤ 2.

Claim: For $I \subseteq [n]$, and $x \in \{0,1\}^{\bar{I}}$, $\widehat{f_{x \to \bar{I}}}(S) = \sum_{T \subseteq \bar{I}} \hat{f}(S \cup T) \chi_T(x)$

Example: $n=4$ $\underbrace{\vdash\!\!+}_{I}\underbrace{\!+\!\!+\!\!\dashv}_{\bar{I}}$.

$$\begin{array}{cc} \pm\hat{f}(0000) & \hat{f}(0100) \quad \cdots \\ \pm\hat{f}(0001) & \vdots \\ \pm\hat{f}(0010) & \vdots \quad \cdots \\ \pm\hat{f}(0011) & \\ \hat{f}_{x\to\bar{I}}(00) & \hat{f}_{x\to\bar{I}}(01) \end{array}$$

If we choose $x \in \{0,1\}^{\bar{I}}$ uniformly, then $\mathbb{E}_x[|\widehat{f_{x\to\bar{I}}}(S)|] \ge |\mathbb{E}_x[\widehat{f_{x\to\bar{I}}}(S)]| = |\hat{f}(S)|$.

Also, $\mathbb{E}_x[\widehat{f_{x\to\bar{I}}}(S)^2] \underset{\text{by claim}}{=} \sum_{T,T' \subseteq \bar{I}} \hat{f}(S\cup T) \cdot \hat{f}(S \cup T') \cdot \mathbb{E}_x[\chi_T(x) \cdot \chi_{T'}(x)] =$

$= \sum_{T \subseteq \bar{I}} \hat{f}(S \cup T)^2$.

Proof: Let (I,x) be a random restriction with $p = \frac{1}{10w}$. By Hastad's switching lemma, $f_{x\to\bar{I}}$ has a decision tree of depth $\le d$ w.p. $\ge 1 - 2^{-d}$.

In such a case, $\sum_{\substack{S \subseteq I \\ |S| > d}} \widehat{f_{x\to\bar{I}}}(S)^2 = 0$. Therefore, $2^{-d} \ge \mathbb{E}_{x,I}\left[\sum_{\substack{S \subseteq I \\ |S|>d}} \widehat{f_{x\to\bar{I}}}(S)^2\right] =$

$= \mathbb{E}_I\left[\sum_{\substack{S \subseteq I \\ |S|>d}} \mathbb{E}_x[\widehat{f_{x\to\bar{I}}}(S)^2]\right] = \mathbb{E}_I\left[\sum_{\substack{S \subseteq I \\ |S|>d}} \sum_{T \subseteq \bar{I}} \hat{f}(S\cup T)^2\right] = \sum_{U \subseteq [n]} \Pr[|U \cap I| > d] \cdot \hat{f}(U)^2$.

For $|U| \ge 20dw$, we get by Chernoff that $\Pr_I[|U \cap I| > d] \ge 1/2$.

(because $|U \cap I|$ is distributed like $\text{Binom}(\ge 20dw, \frac{1}{10w})$).

Hence, $\sum_{|U| \ge 20dw} \hat{f}(U)^2 \le 2^{-d+1}$. ∎

Thm: If f has width-w DNF, then $\sum_U \left(\frac{1}{20w}\right)^{|U|} \cdot |\hat{f}(U)| \le 2$.

Solutions - Homework 2

5. (a). $f(x) = \sum_S \underbrace{\hat f(S)}_{\substack{\geq 0 \\ \Sigma = 1}} \chi_S(x) = \mathop{E}_{S\sim D}[\chi_S(x)]$, where D is the distribution on S given by $\Pr[S] = \hat f(S)$. By Chernoff, for each fixed x, if we choose $S_1, S_2, \dots, S_{cn}$ for some large enough c, then $\left|f(x) - \frac{1}{cn}\sum_{i=1}^{cn} \chi_{S_i}(x)\right| < 0.01$ w.p. $\geq 1 - 2^{-(n+1)}$. Therefore, by union bound, this sample is good for all x simultaneously w.p. $\geq 1/2$. In particular, such sample exists.

(b). Write $f = f^+ - f^-$ where $f^+ = \sum_{S:\, \hat f(S) \geq 0} \hat f(S)\chi_S$ and similarly for f^-. Now apply (a) separately to $\frac{f^+}{\|\hat f^+\|_1}$ and $\frac{f^-}{\|\hat f^-\|_1}$ with accuracies $\frac{0.01}{2\|\hat f^+\|_1}$ and $\frac{0.01}{2\|\hat f^-\|_1}$.

1. Consider $f = \chi_{\{1,2,3\}} = \chi_{\{1\}} \cdot \chi_{\{2\}} \cdot \chi_{\{3\}}$. f is $1/2$-far from dictators. We claim that the test must accept f w.p. 1: Assume the test checks that

$f(x)\cdot f(y) \cdot f(z) = -1$; $f(x)f(y)f(z) = \chi_{\{1\}}(x)\chi_{\{2\}}(x)\chi_{\{3\}}(x)\cdot \ldots \cdot \chi_{\{3\}}(z) = (-1)(-1)(-1) = -1$.

So test must accept f. Similarly, for $f(x)\cdot f(y)\cdot f(z) = 1$.

7. 

(a). If $val(G) \geq 1-\lambda$ then consider the assignment $f_v = \chi_{\{L(v)\}}$, where L is the assignment to the unique label cover. W.p. $\geq 1-\lambda$, the constraint $\{u,v\} \in E$ chosen by the tester is s.t. $\sigma_{u\to v}(L(u)) = L(v)$. In such a case the tester accepts w.p. $1-\delta$. So overall acceptance prob. is $(1-\lambda)(1-\delta) \geq 1 - \lambda\delta$.

(b). Consider the assignment giving some dictator to each f_v. Then the test applies $Hastad_\delta$ to the average of two dictators, $\chi_{\{i\}}$ and $\chi_{\{j\}}$.

If $i=j$ then the test accepts w.p. $1-\delta$.

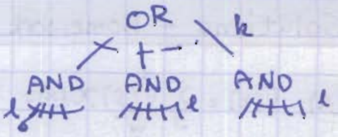
If $i \neq j$ the test accepts w.p. $\frac{5}{8} - \frac{\delta}{4}$ (since the acceptance prob. is $\frac12 + \frac12 \sum_S (1-2\delta)^{|S|} \hat f(S)^3$)

2. (a). A function is 1-resilient if it is balanced and remains balanced after fixing any one coordinate.

(b). Consider the test that takes x uniformly and $w \sim \mu_\delta$ for $\delta = \frac{1-\rho}{2}$ and checks that $f(x)\cdot f(x+w) = 1$. Its acceptance prob. is, $\frac12 + \frac12 \sum_S \rho^{|S|} \cdot \hat f(S)^2$

If f is 1-resilient then this is $\leq \frac12 + \frac12 \rho^2$. If $\exists S.\ |S| \leq 1$ and $|\hat f(S)| \geq \varepsilon$ then this is $\geq \frac12 + \frac12 \rho \varepsilon^2$. Choosing $\rho = \frac{\varepsilon^2}{2}$ makes the first case $\leq \frac12 + \frac18 \varepsilon^4$ and the second $\geq \frac12 + \frac14 \varepsilon^4$. We can distinguish the two with 90% confidence using $\Theta(1/\varepsilon^8)$ repetitions.

3. (a). $2^{-(\ell-1)}\cdot(1-2^{-\ell})^{k-1}$



(b). $(1-2^{-\ell})^k = \Pr[\text{Tribes}=0]$. So for any ℓ we can choose $k = \lfloor 2^{\ell}\cdot\ln 2\rfloor$.

(c). All influences are $\Theta(2^{-\ell}) = \Theta\left(\frac{\log n}{n}\right)$ where $n = k\cdot\ell$. For majority this is $\Theta\left(\frac{1}{\sqrt{n}}\right)$.

[KKL] showed that this is best possible.

4. (a). $E[f(x)\cdot g(x)] - E[f]\cdot E[g] = \langle f,g\rangle - \hat{f}(\phi)\hat{g}(\phi) = \sum_{S\neq\phi}\hat{f}(S)\hat{g}(S)$.

(b). Assume f depends on $\{1,..,r\}$. $Cov[h,f] = \sum_{S\neq\phi}\hat{f}(S)\hat{h}(S) = \sum_{[r]\supseteq S\neq\phi}\hat{f}(S)\hat{h}(S) \le$

$\le \sqrt{\sum_{\substack{S\neq\phi\\ S\subseteq[r]}}\hat{f}(S)^2}\cdot\sqrt{\sum_{\substack{S\neq\phi\\ S\subseteq[r]}}\hat{h}(S)^2} \le 1\cdot\sqrt{\sum_{\substack{S\neq\phi\\ S\subseteq[r]}}\hat{h}(S)^2}$. By assumption, $\forall i.\ \sum_{S\ni i}(1-\delta)^{|S|}\hat{h}(S)^2 \le \varepsilon$.

Therefore, $(1-\delta)^r\cdot\sum_{\substack{S\neq\phi\\ S\subseteq[r]}}\hat{h}(S)^2 \le \sum_{i=1}^{r}\sum_{S\ni i}(1-\delta)^{|S|}\cdot\hat{h}(S)^2 \le \varepsilon\cdot r$.

Possible: $\min(1-\delta, (1-\delta)^r\cdot r)$ $\le |S|(1-\delta)^{|S|}$ $(1\le|S|\le r)$

6. (a). ... $\|f^{odd}\|_2^2 = \sum_{|S|\,odd}\hat{f}(S)^2 \le \sum_{S\neq\phi}\hat{f}(S)^2 \le \sum_{S}|S|\hat{f}(S)^2 = I(f)$.

(b). Write $F = (f_1, f_2, \dots, f_m)$ and apply (a) to each f_i and sum the ineq.

(c). Assume F is an embedding with distortion D.

$$n^2 \le E_x\left[\|F(x) - F(x\oplus(1,..,1))\|_2^2\right] \le \sum_{i=1}^{n} E_x\left[\|F(x) - F(x\oplus e_i)\|_2^2\right] \le n\cdot D^2 \Rightarrow D \ge \sqrt{n}. \blacksquare$$

13.3.2008 Observations: 1. $\widehat{f_{x\to\bar{I}}}(S) = \sum_{T\subseteq\bar{I}}\hat{f}(S\cup T)\cdot\chi_T(x)$

2. $E_x\left[|\widehat{f_{x\to\bar{I}}}(S)|\right] \ge |\hat{f}(S)|$

3. $E_x\left[\widehat{f_{x\to\bar{I}}}(S)^2\right] = \sum_{T\subseteq\bar{I}}\hat{f}(S\cup T)^2$.

Thm: If f is computable by a width-w DNF, then $\forall d\ge 5$, $\sum_{|S|\ge 20dw}\hat{f}(S)^2 \le 2^{-d+1}$.

Remark: This gives $n^{O(w\log 1/\varepsilon)}$ time alg. to learn DNFs from random examples.

Also $n^{O(\log n\cdot\log 1/\varepsilon)}$ for poly-size DNFs.

Thm: If f has width-w DNF then $\sum_{U}\left(\frac{1}{20w}\right)^{|U|}\cdot|\hat{f}(U)| \le 8$.

Proof: Let (I,x) be a random restriction with $p = \frac{1}{20w}$. Then,

$$E_{I,x}\left[\|\widehat{f_{x\to\bar{I}}}\|_1\right] \underset{\|\cdot\|_1 \text{ of depth-}d \text{ DT is} \le 2^d}{\le} \sum_{d=0}^{\infty}\Pr[\text{DT-depth}(f_{x\to\bar{I}}) = d]\cdot 2^d \underset{\text{Hastad}}{\le} \sum_{d=0}^{\infty}4^{-(d-1)}\cdot 2^d = 8.$$

On the other hand, $E_{I,x}\left[\|\widehat{f_{x\to\bar{I}}}\|_1\right] = E_I\left[\sum_{S\subseteq I}E_x\left[|\widehat{f_{x\to\bar{I}}}(S)|\right]\right] \underset{\text{by (2)}}{\ge} E_I\left[\sum_{S\subseteq I}|\hat{f}(S)|\right] =$

$= \sum_{U\subseteq[n]}\Pr_I[U\subseteq I]\cdot|\hat{f}(U)| = \sum_{U\subseteq[n]}\left(\frac{1}{20w}\right)^{|U|}\cdot|\hat{f}(U)|$. $\blacksquare$

This implies that $\sum_{|U|\le O(w\log 1/\varepsilon)}|\hat{f}(U)| \le w^{O(w\log 1/\varepsilon)}$.

Cor: If f is computable by a width w DNF, then it is ε-concentrated on a set of size $w^{O(w \log 1/\varepsilon)}$.

Proof: Define $S = \{U : |U| \le O(w \log 1/\varepsilon) \text{ and } |\hat{f}(U)| \ge \frac{\varepsilon}{w^{O(w\log 1/\varepsilon)}}\}$. Then $|S| \le w^{O(w \log 1/\varepsilon)}$.

Moreover, $$\sum_{U \notin S} \hat{f}(U)^2 \le \underbrace{\sum_{|U| \ge O(w\log 1/\varepsilon)} \hat{f}(U)^2}_{\le \varepsilon} + \underbrace{\sum_{\substack{|U| \le O(w \log 1/\varepsilon) \\ |\hat{f}(U)| \le \frac{\varepsilon}{w^{O(w\log 1/\varepsilon)}}}} \hat{f}(U)^2}_{\le \varepsilon \text{ since } \sum \hat{f}(U)^2 \le \max |\hat{f}(U)| \cdot \sum |\hat{f}(U)|} \le 2\varepsilon. \quad \blacksquare$$

This gives: $w^{O(w \log 1/\varepsilon)}$ learning time algorithm for width-w DNF and $n^{O(\log \log n \cdot \log 1/\varepsilon)}$-time alg. for poly-size DNF [Mansour].

Summary: (uniform PAC, constant ε)

	random examples	queries
poly-size depth d circuits	$n^{O(\log^{d-1} n)}$ [LMN93]	$\rightarrow$ ...
poly-size DNF	$n^{O(\log n)}$ -"-	poly(n) [Jackson94]
poly-size DT	$n^{O(\log n)}$ -"-	poly(n)
log n-junta	$n^{0.704 \log n}$	poly(n)

Open problems:

- Are poly(n) size DNFs concentrated on poly(n) coefficients? (this would imply Jackson). Tribes may be counterexample?!
- Big open question: can monotone polysize DNFs be learned from random examples in polytime? (CCC'06: O'Donnell & Servedio did this for monotone DTs).

Learning Juntas (Mossell, O'Donnell, Servedio, STOC'03)

We want to learn k-juntas. Think of k as being very small, say, $k = \log n$.

Because a k-junta has $\le 2^k$ nonzero Fourier coeff all in levels $\le k$ and multiple of 2^{-k}, we can learn them exactly in time $\mathrm{poly}(2^k, n)$ using queries, or in time $n^k \cdot \mathrm{poly}(2^k, n)$ using random examples. We will improve this slightly to $n^{0.704k} \cdot \mathrm{poly}(2^k, n)$.

Step 1: Finding a relevant coordinate is enough.

Prop: If there is an algorithm for finding a relevant coordinate in a given (nonconstant) k-junta in time $n^{\alpha(k)} \cdot \mathrm{poly}(2^k, n)$ using random examples then there is an algorithm for learning k-juntas in the same time.

Proof sketch: Find a relevant coordinate, say i. Now recur on $f_{0\to i}$ and on $f_{1\to i}$. Notice that we can simulate examples from restrictions of f of r variables using examples from f with an overhead of 2^r (on average). After finding all k relevant coordinates we can determine the junta by observing examples for all 2^k possible settings. This takes $\Theta(k\cdot 2^k)$ examples. ∎

If a k-junta has a nonzero Fourier coeff at level $d>0$ then we can find it in time $n^d \cdot \mathrm{poly}(n, 2^k)$. This gives a relevant coordinate.

Does a junta always have a nonzero Fourier coeff. in level $0<d<k$?

No! For instance, parity $\chi_{\{i_1, i_2, \ldots, i_k\}}$. But parities are easy to learn!

Given examples $(x^1, f(x^1)), (x^2, f(x^2)), \ldots$ write a sequence of linear equalities over GF_2 in variables $a_1, a_2, \ldots, a_n$:

$$a_1 x_1^1 + a_2 x_2^1 + \ldots + a_n x_n^1 = f(x^1)$$
$$a_1 x_1^2 + a_2 x_2^2 + \ldots + a_n x_n^2 = f(x^2)$$
$$\vdots$$

We only need $\sim n$ equations.

Step 2: Multilinear polynomials over GF_2.

Examples: $\mathrm{PARITY}(x_1, \ldots, x_n) = x_1 + x_2 + \ldots + x_n$ of degree 1.

$$\mathrm{AND}(x_1, \ldots, x_n) = x_1 \cdot x_2 \cdot \ldots \cdot x_n \quad \text{of degree } n.$$

Prop: Any $f: GF_2^n \to GF_2$ can be uniquely expressed as a multilinear polynomial.

Proof: We can write $f(x) = \sum_{a \in GF_2^n} f(a) \cdot \prod_{i=1}^{n} (1 - (x_i - a_i))$. Uniqueness follows because the only multilin. poly. that is constantly 0 is the zero poly. ∎

Thm: The class $\{f: GF_2^n \to GF_2 \mid \deg_{GF_2}(f) \le e\}$ is learnable from random examples in time $n^{\omega e} \cdot \mathrm{poly}(n, 2^e)$ where ω is the matrix multiplication constant (currently $\omega \le 2.376$ [Coppersmith Winograd 90], Strassen $\log_2 7$).

Proof: Take $m \approx n^e$ random examples (requires proof) $(x^1, f(x^1)), (x^2, f(x^2)), \ldots$ and solve equations in variables $\{a_T\}_{|T| \le e}$ over GF_2:

$$\Big\{ \sum_{|T| \le e} a_T \cdot \underbrace{\prod_{j \in T} x_j^i}_{\text{const!}} = f(x^i) \Big\}_{i=1}^{m}$$

This takes time $n^{\omega e} \cdot \mathrm{poly}(n, 2^e)$.

Thm [Siegenthaler 84']: If $f: \{0,1\}^k \to \{-1,1\}$ is s.t. $\hat{f}(S) = 0$ for all S with $1 \le |S| \le d$ then over GF_2, we can express f as a degree $\le k-d$ multilinear poly.

Proof: Let $g = f \cdot \chi_{\{1,\ldots,k\}}$. If we can represent g as a degree $\le k-d$ multilin. poly over GF_2 then we're done because we obtain f by adding $x_1 + x_2 + \ldots + x_n$ which is of degree 1. Notice that $\forall S$, $\hat{g}(S) = \hat{f}(S \Delta \{1, 2, \ldots, k\})$, hence $\hat{g}(S) = 0$ for all S with

$k-d \le |S| \le k-1$. Consider the multilinear poly. over $\mathbb{R}$, $h(x_1,\dots,x_n) = \frac{1}{2} - \frac{1}{2}\sum_S \hat{g}(S)\prod_{i\in S}(1-2x_i)$.

Clearly, for all $x_1,\dots,x_n \in \{0,1\}$, $h(x_1,\dots,x_n) = \frac{1}{2} - \frac{1}{2}g(x_1,\dots,x_n) \in \{0,1\}$

This implies that if we expand h, then the coeff of every monomial is integer (this can be seen by induction on the degree of the monomial: the free coeff. is $h(0,\dots,0)$ and hence integer; the x_i coeff is $h(0,\dots,1,0,\dots,0) - h(0,\dots,0) \in \mathbb{Z}$, and is therefore integer). If we reduce each coeff of h modulo 2, we get a representation of g as a multilin. poly over GF_2. It remains to prove that the degree is $\le k-d$.

Notice that $\prod_{i\in S}(1-2x_i)$ is of degree $=|S|$. This means that the coeff of $\prod_{i\in T} x_i$ for some T of size $\geq k-d$ is only affected by $\hat{g}(\{1,\dots,k\})$ and is therefore $-\frac{1}{2}\hat{g}(\{1,\dots,k\})\cdot(-2)^{|T|}$.

Since all coeff of h are integers, $-\frac{1}{2}\hat{g}(\{1,\dots,k\})(-2)^{k-d}$ must be integer. This implies that for T of size $\geq k-d+1$ the coeff is even and therefore disappears when we take modulo 2. ▨

<u>Remark</u>: We could avoid the last complication by using the homework (unbalanced functions have "low" Fourier coeff).

<u>Thm</u>: k-juntas can be learned in time $n^{\omega k/\omega+1}\cdot \mathrm{poly}(n,2^k)$, from random examples.

<u>Proof</u>: Let $d = \frac{\omega k}{\omega+1}$ $(\approx 0.7k)$. Look for nonzero Fourier coeff up to level d (in levels $1,\dots,d$) in time $n^d\cdot\mathrm{poly}(n,2^k)$. If found, we have a relevant coordinate, and we're done. Otherwise, our function is of degree $\le k-d = \frac{k}{\omega+1}$ over GF_2, so we can learn it in time $n^{\frac{\omega k}{\omega+1}}\cdot\mathrm{poly}(n,2^k)$. ▨

<u>Open Questions</u>:
- Learn k-juntas in time $\mathrm{poly}(n,2^k)$ or even $n^{k/2}\cdot\mathrm{poly}(n,2^k)$.
- Do something similar over $\{0,1,2\}^n$, or over $\{0,1\}^n$ with μ_p.