Today:

- Learning via Fourier Coeffs. (cont.)
  - Finish low degree alg
  - applications
- Possibly - learning heavy Fourier Coeffs (even if "high degree")

Recall from last time:

- Can approx any single fourier Coeff (*) from random samples:

  $\left.\begin{array}{l}\text{additive error } \gamma \\ \text{confidence error } \delta\end{array}\right\}$ use $O\left(\frac{1}{\gamma^2}\log\frac{1}{\delta}\right)$ Samples

- <u>def.</u> $f: \{\pm 1\}^n \to \mathbb{R}$ has $\alpha(\varepsilon, n)$-Fourier concentration if $\sum_{\substack{S \text{ s.t.} \\ |S| > \alpha(\varepsilon,n)}} \hat{f}(S)^2 \leq \varepsilon \quad \forall\, 0 \leq \varepsilon \leq 1$

  (For Boolean f, equivalent to $\sum_{\substack{S \text{ s.t.} \\ |S| \leq \alpha(\varepsilon,n)}} \hat{f}(S)^2 \geq 1-\varepsilon$)

Low degree Algorithm

Given d (degree) τ (accuracy) δ (confidence)

Algorithm:

$n^d \approx$ # F.C. of deg $\leq d$

- Take $m = O\left(\frac{n^d}{\tau} \ln \frac{n^d}{\delta}\right)$ samples
- $C_S \leftarrow$ estimate of $\hat{f}(S)$ $\forall S$ s.t. $|S| \leq d$
- $h(x) \equiv \sum_{|S|<d} C_S \chi_S(x)$
- Output function $\underbrace{\text{sign}}_{\text{corrects } h \text{ to make it Boolean}}(h(\cdot))$

Two stages:

today

1) f has low Fourier conc $\Rightarrow \underbrace{E_x[(f(x)-h(x))^2]}_{\text{small}}$ ✓

2) $\underbrace{Pr_x[f(x) \neq \overbrace{\text{sign}(h(x))}^{\text{hypothesis}}]}_{\text{error of hypothesis we output}} \leq E_x[(f(x)-h(x))^2]$ ✓

last time

Thm if f has $d = \alpha(\varepsilon, n)$ - F.C. then h satisfies $E_x[(f(x)-h(x))^2] \le \underbrace{\varepsilon}_{\text{F.C.}} + \underbrace{\tau}_{\text{approx error in alg est of } C_S\text{'s}}$

Pf.

Claim with prob $\ge 1-\delta$, $\forall S$ s.t. $|S| \le d$ $|C_S - \hat{f}(S)| \le \gamma$ for $\gamma \leftarrow \sqrt{\frac{\tau}{n^d}}$

Pf. (union bnd + *) note: $\frac{1}{\gamma^2} = \frac{n^d}{\tau}$ & $\tau = n^d \cdot \gamma^2$

$O(\frac{n^d}{\tau} \ln \frac{n^d}{\delta}) = O(\frac{1}{\gamma^2} \ln \frac{n^2}{\delta})$ samples

yields $\Pr[|C_S - \hat{f}(S)| \ge \gamma] < \frac{\delta}{n^d}$ $\forall S$

union bnd $\Rightarrow \Pr[\exists S \text{ s.t. } |C_S - \hat{f}(S)| > \gamma] < \delta$

since only $\le n^d$ S's of deg $\le d$

(Claim)

Assume $\forall S$ s.t. $|S| \le d$ $|C_S - \hat{f}(S)| \le \gamma$

define $g(x) \equiv f(x) - h(x)$ "error of h"

Fourier transform is linear $\Rightarrow$ $\forall s\ \hat{g}(s) = \hat{f}(s) - \hat{h}(s)$

by defn $\quad \forall s$ s.t. $|s| > d \quad \hat{h}(s) = 0$

$\Rightarrow \hat{g}(s) = \hat{f}(s)$

$\forall s$ s.t. $|s| \leq d$

$\Rightarrow \hat{g}(s) = \hat{f}(s) - C_s$

so $\hat{g}(s)^2 \leq \gamma^2$

$$\mathop{E}_{x}\left[\left(f(x) - h(x)\right)^2\right] = E\left[g(x)^2\right]$$

$$= \sum_{s} \hat{g}(s)^2 \qquad \text{Parseval}$$

$$= \sum_{\substack{s \text{ s.t.} \\ |s| \leq d}} \underbrace{\hat{g}(s)^2}_{\leq \gamma^2} + \underbrace{\sum_{|s| > d} \underbrace{\hat{g}(s)^2}_{\hat{f}(s)^2}}_{\leq \varepsilon \text{ by F.C of } f}$$

$$\leq \underbrace{\binom{n}{d} \cdot \gamma^2}_{= \tau} + \varepsilon$$

$$\leq \underbrace{\tau}_{\text{sampling error}} + \varepsilon$$

inherent error in approx via low degree

∎

today **Thm** f has $d = \alpha(\varepsilon, n)$ - F.C.

$\Rightarrow h$ satisfies $E_x[(f(x)-h(x))^2] < \varepsilon + \tau$

with prob $\geq 1-\delta$

last time **Thm** $f : \{\pm 1\}^n \to \{\pm 1\}$

$h : \{\pm 1\}^n \to \mathbb{R}$

then $\Pr[f(x) \neq \mathrm{sign}(h(x))] \leq E[(f(x)-h(x))^2]$

Correctness of learning alg:

Thm if $\mathcal{C}$ is class of fctns

with Four. conc. $d = \alpha(\varepsilon, n)$

then there is a $q = O\left(\frac{n^d}{\varepsilon} \log \frac{n^d}{\delta}\right)$ samples

(from uniform dist) learning alg for $\mathcal{C}$

Pf run low deg with $\tau = \varepsilon$

get h with error $\leq 2\varepsilon$

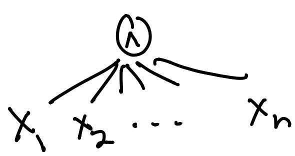
Applications:

1) bounded depth decision trees
2) "and" fctn
3) fctns of few vars "Junta fctns"
   (small # relevant vars)
4) AC(0) fctns: (const depth, poly gates)

Consider constant depth ckts:

<u>def</u> Boolean ckt C is a DAG

gates: $\underbrace{\wedge}_{\text{and}}, \underbrace{\vee}_{\text{or}}, \underbrace{\neg}_{\text{not}}, 1, 0, \underbrace{x_1 \dots x_n}_{\text{vars}}$



and fctn.

1 gate with n inputs

What can we compute in const depth?

<u>everything</u> e.g. Karnaugh maps

$\Rightarrow$ depth 2 circuit for any f.

<u>But</u> exponential size

Can we compute parity in const depth? $AC(0)$
with poly # gates?

NO [Furst Saxe Sipser]

lemon

lemons $\Rightarrow$ lemonade:

proofs of parity $\notin AC(0)$
$\Rightarrow$ following:

Thm [Hastad Linial Mansour Nisan]

$\forall f$ computable via size s depth d ckts

$$\sum_{|S|>t} \hat{f}^2(S) \le \alpha \quad \text{for} \quad t = O\left(14 \log \frac{2s}{\alpha}\right)^{d-1}$$

take $\left.\begin{array}{l} s = \text{poly}(n) \\ d = \text{const} \\ \alpha = O(\varepsilon) \end{array}\right\}$ deg $t = O\left(\log^d\left(\frac{n}{\varepsilon}\right)\right)$

$\Rightarrow$ $n^{O\left(\log^d\left(\frac{n}{\varepsilon}\right)\right)}$ sample algorithm
to learn any $f \in AC0$

($n^{O(\log\log n)}$ sample algorithms are known)

5) linear threshold fctns

(plus a little more work)

New topic (in Fourier):

learning "large" Fourier coeffs.

arises in learning theory, complexity theory, coding theory, Compressive sensing.

<u>PAC Setting</u>:

from uniform for today

Given samples $x, f(x)$

Find χ_S s.t. χ_S & f agree $\underbrace{\text{"a lot"}}_{\text{large F.C.}}$

Thought to be hard:

if x from arbitrary dist then NP-hard

" " " unif then still thought to be hard in terms of time

"hardness of parity with noise"
"hardness of decoding linear codes"

hardness assumption used for crypto
learning

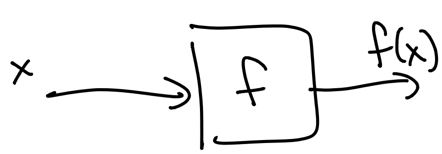
If noise random:

"hardness of decoding random linear codes"
"noisy parity"

[A. Blum, Kalai, Wasserman]
can solve in $2^{O(n/\log n)}$

When $x, f(x)$ arrive as random samples
then computationally hard

What if we get query access
to $x, f(x)$?



black box

Given f, θ

output all close fctns

1) output all coeff S s.t. $|\hat{f}(S)| \geq \theta$

2) only output S s.t. $|\hat{f}(S)| \geq \frac{\theta}{2}$

only output close fctns (no real junk)

in between case:

$\frac{\theta}{2} \leq |\hat{f}(S)| \leq \theta$

ok to output

ok not to output

Concern if a lot of S satisfy 1

output size is huge?

Boolean Parseval's: $\sum \hat{f}^2(S) = 1$

$$1 \overset{B.P.}{=} \sum_{\text{all } S} \hat{f}^2(S) \geq \underbrace{\sum_{\substack{S \text{ s.t.} \\ |\hat{f}(S)| \geq \frac{\theta}{2}}} \hat{f}^2(S)}_{\text{allowed to output}} \geq \underbrace{\left(\#S \text{ s.t. } \hat{f}(S) \geq \frac{\theta}{2}\right)}_{M} \cdot \frac{\theta^2}{4}$$

$M \leq \frac{\cancel{4}}{\theta^2}$ max # of S s.t. $\hat{f}(S) \geq \frac{\theta}{2}$

M is $O\left(\frac{1}{\theta^2}\right)$

recall $\Pr_x[f(x) = \chi_S(x)] = \frac{1}{2} + \frac{\hat{f}(S)}{2}$

So case 1 $\Rightarrow \Pr_x[f(x) = \chi_S(x)] \geq \frac{1}{2} + \frac{\theta}{2}$

2 $\Rightarrow$ $\leq \frac{1}{2} + \frac{\theta}{4}$

Warmup:

① poly queries unbounded time $\Rightarrow$ find all $\hat{f}(S)$'s

② want $\Pr_x[f(x) = \chi_S(x)] = 1$ i.e. $f(x) \equiv \chi_S(x)$

"no noise" i.e. $\hat{f}(S) = 1$ ← only one F.C. is nonzero

Algorithm: 1) take a bunch of samples & solve lin system to find coefficients

2) $\forall i \in [n]$

figure out if $i \in S$

put i in S if $f(1\ldots1) \neq f(\underbrace{1\ldots1(-1)1\ldots1}_{e_i})$ ← -1 in i^{th} locn

Output S

$e_i = (1111(-1)1111)$ ← i^{th} locn

$\chi_S(x) = \prod_{i \in S} x_i$ $\qquad$ $\chi_S(1 \dots 1) = \prod_{i \in S} 1 = 1$

$$\chi_S(1 \dots \underset{i}{(-1)} \dots 1) = \prod_{j \in S \setminus \{i\}} 1 \cdot (-1)$$

$$= -1$$

③ Suppose f agree with χ_S on $\geq \frac{7}{8}$ for some S

(here BP $\Rightarrow$ ≤ 3 such S can be output, but actually ≤ 1 such S can be output)

Algorithm:

choose $r_1 \dots r_t$;

$\forall i \in [n]$:

put i in S if for majority of r_j's $f(r_j) \neq f(\underbrace{r_j \oplus e_i}_{\text{flip } i^{\text{th}} \text{ bit in } r_j})$

Output S

Why does it work?

self-correcting linear fctns

$$\Pr[f(\overset{\swarrow \text{uniform}}{r_j}) \text{ disagrees with } \chi_S(r_j)] \le \frac{1}{8}$$

$$\Pr[f(\underbrace{r_j \oplus e_i}_{\text{uniform}}) \quad " \quad " \quad \chi_S(r_j \oplus e_j)] \le \frac{1}{8}$$

$\Rightarrow \Pr[\text{ both agree}] \ge 3/4$

So if both agree & if $i \in S$

then $\chi_S(r_j) \ne \chi_S(r_j \oplus e_i)$

so $f(r_j) \ne f(r_j \oplus e_i)$