Lecture 15

Fourier-based learning algorithms

- learning one Fourier coeff
- the low degree algorithm
- Fourier concentration

Review from last time:

def uniform distribution learning algorithm for concept class $\mathcal{C}$ is algorithm $\mathcal{A}$ s.t.

- $\mathcal{A}$ given $\varepsilon, \delta > 0$ access to $Ex(f)$ for $f \in \mathcal{C}$
- $\mathcal{A}$ outputs h s.t. with prob $\geq 1-\delta$ error(h) w.r.t. f is $\leq \varepsilon$

(h is ε-close to f)

Parameters of interest:

- m — # samples used by $\mathcal{A}$ "sample complexity"
- ε — accuracy parameter
- δ — confidence parameter
- runtime? — hope for poly$(\log(\text{domainsize}), \frac{1}{\varepsilon}, \frac{1}{\delta})$

- description of h?
  - should it be similar to description of fctns in $\mathcal{C}$? "proper learning"
  - at least should be relatively compact + efficient to evaluate
    - compact: $O(\log |\mathcal{C}|)$

Fourier Representation

$S \subseteq \{1..n\}$

parity fctns

for $x \in \{\pm 1\}^n$, $\chi_S(x) = \prod_{i \in S} x_i$

define $\langle f, g \rangle = \frac{1}{2^n} \sum_{x \in \{\pm 1\}^n} f(x)\, g(x)$ inner product (but normalized)

Def. $\hat{f}(S) \equiv \langle f, \chi_S \rangle$

$= \frac{1}{2^n} \sum_{x \in \{\pm 1\}^n} f(x) \chi_S(x)$

Fourier Coefficients of f

Thm $\forall f \quad f(x) = \sum \hat{f}(S) \chi_S(x)$

Parseval's identity:

$\langle f, f \rangle = \sum_S \hat{f}(S)^2$

Thm $\hat{f}(S) = 1 - 2 \Pr[f(x) \neq \chi_S(x)]$

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

Claim: if f doesn't "depend" on x_j then $\forall S$ s.t. $j \in S$, $\hat{f}(S) = 0$

Learning via Fourier Representation

will look at learning algorithms that are based on estimating Fourier representation of fctn f (similar to polynomial interpolation)

Approximating one Fourier coefficient:

lemma for any $S \subseteq [n]$, can approx $\hat{f}(S)$ to within additive γ (i.e. $|\text{output} - \hat{f}(S)| \leq \gamma$) with prob $\geq 1-\delta$ in $O(\frac{1}{\gamma^2} \log \frac{1}{\delta})$ samples.

no queries needed!

Pf. Chernoff + $\hat{f}(S) = 2 \Pr_x [f(x) = \chi_S(x)] - 1$

estimate this

■

Can we find any or all heavy coefficients?

there are exponentially many coeffs.

Can use same samples to estimate each coeff, but must union bound prob of error (error = bad approx) on any of them.

Need $\delta \ll \frac{1}{2^n}$, which needs $O(\frac{1}{\gamma^2} \cdot n)$ samples, but exponential runtime. ← turns out that queries help a lot

What if we "know where to look" for heavy coeffs?

e.g. all heavy coeffs are in "low degree" coeffs? If so, can search!

END REVIEW

Fourier Representations of Important Examples

Recall:
$-1 \leftrightarrow T$
$+1 \leftrightarrow F$

1) AND on $T \subseteq N$ s.t. $|T| = k$

$$AND(x) = \begin{cases} -1 & \text{if } \forall i_j \in T = \{i_1 \dots i_k\} \\ & \quad x_{i_j} = -1 \\ +1 & \text{o.w.} \end{cases}$$

$$f(x) = \begin{cases} 1 & \text{if } \forall i \in T \; x_i = -1 \\ 0 & \text{o.w.} \end{cases}$$

} AND w/output over $\{0,1\}$ range

$$= \frac{(1-x_{i_1})}{2} \cdot \frac{(1-x_{i_2})}{2} \cdots \frac{(1-x_{i_k})}{2}$$

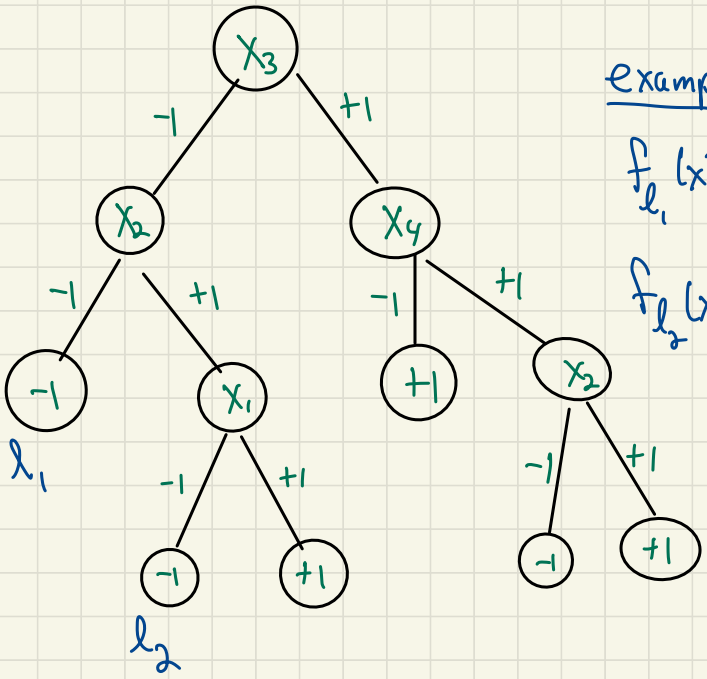
$$= \sum_{S \subseteq T} \frac{(-1)^{|S|}}{2^k} \chi_S$$

$$AND = 1 - 2f(x) = 1 - \frac{2}{2^{|T|}} - \sum_{\substack{S \subseteq T \\ S \neq \emptyset}} \frac{(-1)^{|S|}}{2^{k-1}} \chi_S$$

$0 \to -1$
$1 \to +1$

Note: all Fourier coeffs containing vars not in T are 0

2) Decision trees



examples

$$f_{\ell_1}(x) = \frac{(1-x_3)}{2} \cdot \frac{(1-x_2)}{2}$$

$$f_{\ell_2}(x) = \frac{(1-x_3)}{2} \cdot \frac{(1+x_2)}{2} \cdot \frac{(1-x_1)}{2}$$

First consider path functions:

$$f_\ell(x) = \prod_{i \in V_\ell} \frac{(1 \pm x_i)}{2}$$

(left or right)

(V_ℓ: vars visited on path to leaf ℓ)

$$= \frac{1}{2^{|V_\ell|}} \sum_{S \subseteq V_\ell} (\pm 1) \chi_S$$

($(-1)^{\#\text{ left turns taken in } S}$)

$$= \begin{cases} 1 & \text{if } x \text{ takes path to } \ell \\ 0 & \text{o.w.} \end{cases}$$

exactly one of these is 1. all others are 0.

So $f(x) = \sum_{\ell \in \text{leaves of } T} f_\ell(x) \cdot \text{val}(\ell)$

Comment only coeffs corresponding to S st. $|S| \leq$ max path length have a hope of being non-zero.

The low degree algorithm

definition of fctns for which low degree Fourier coeffs pretty much suffice to describe fctn.

def $f: \{\pm 1\}^n \to \mathbb{R}$ has $\alpha(\varepsilon, n)$ - Fourier concentration if

$$\sum_{\substack{S \subseteq [n] \\ \text{s.t.} \\ |S| > \alpha(\varepsilon,n)}} \hat{f}(S)^2 \leq \varepsilon \qquad \forall\ 0 < \varepsilon < 1$$

↑ for Boolean f, this implies

$$\sum_{\substack{S \subseteq [n] \\ \text{s.t.} \\ |S| \leq \alpha(\varepsilon,n)}} \hat{f}(S)^2 \geq 1 - \varepsilon$$

examples

1) fctn f which depends on $\leq k$ vars has $\sum_{S \text{ s.t. } |S| > k} \hat{f}(S)^2 = 0$

} if f doesn't depend on x_i then all $\hat{f}(S)$ for which $i \in S$ satisfy $\hat{f}(S) = 0$

2) $f = \text{AND}$ on $T \subseteq \{1..n\}$ has $\log(\frac{4}{\varepsilon})$-F.C.

- all $\hat{f}(S)^2 = 0$ for $|S| > |T|$ (pairing argument)
- if $|T| \leq \log \frac{4}{\varepsilon}$ then ✓
- if $|T| \geq \log \frac{4}{\varepsilon}$ then: (almost always False)

$$\hat{f}(\varphi)^2 = (1 - 2\Pr(f(x) \neq \chi_\varphi(x)))^2$$

empty set (constant fctn)

$$= \left(1 - \frac{2}{2^{|T|}}\right)^2$$

$$> 1 - \varepsilon$$

So $\sum_{S \neq \varphi} \hat{f}(S)^2 \leq \varepsilon$ & f has 0-F.C.

Idea: Can we approximate f by only considering low degree Fourier Coeffs?

Low degree algorithm

approximates fctns with $d \equiv \alpha(\varepsilon, n)$ Fourier concentration.

Given:
- d degree
- γ accuracy (will set $\gamma := \varepsilon$)
- δ confidence

Algorithm:

- Take $m = O\left(\frac{n^d}{\gamma} \ln \frac{n^d}{\delta}\right)$ samples
- For each S s.t. $|S| \leq d$: ($\leq \binom{n}{d}$ of these; reuse samples)

  $C_S \leftarrow$ estimate of $\hat{f}(S)$
- let $h(x) \equiv \sum_{|S| \leq d} C_S \cdot \chi_S(x)$
- output $\operatorname{sign}(h)$ as hypothesis

Why does this work?

Two stages:

1) Show that f has low F.C.

$$\Rightarrow \mathbb{E}_x[(f(x)-h(x))^2] \text{ small}$$

L_2-dist $= \frac{\text{dist}}{2^n}$

2) Show that $\Pr[f(x) \neq \text{sign}(h(x))] \leq \mathbb{E}_x[(f(x)-h(x))^2]$

↑ Hamming dist

put together:
f has low F.C. $\Rightarrow$ sign$(h(x))$ is good approximation of f

First "stage":

Thm 1 if f has $d = \alpha(\varepsilon, n)$ - F.C. then h satisfies $\mathbb{E}_x[(f(x)-h(x))^2 \leq \varepsilon + \tau$ with prob $\geq 1-\delta$

Pf (1) each low degree Fourier coeff is well approximated:

Claim with prob $\geq 1-\delta$, $\forall S$ s.t. $|S| \leq d$

$$|C_S - \hat{f}(S)| \leq \gamma \quad \text{for} \quad \gamma \leftarrow \sqrt{\frac{\tau}{n^d}}$$

Pf of claim (Chernoff + union bnd)

note, $\frac{1}{\gamma^2} = \frac{n^d}{\tau}$

Chernoff bnd $\Rightarrow$

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

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

union bnd over all $\binom{n}{d}$ S's $\Rightarrow$

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

■

(2) all low degree Fourier coeffs well approx $\Rightarrow$ low l_2 error:

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

define $g(x) \equiv f(x) - h(x)$

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

by defn, $\forall S$ s.t. $|S| > d$, $\hat{h}(S) = 0 \Rightarrow \hat{g}(S) = \hat{f}(S)$

$|S| \le d$, $\hat{h}(S) = C_S$

$\Rightarrow$ $\hat{g}(S) = \hat{f}(S) - C_S$

so $\hat{g}(S)^2 \le \gamma^2$

So $E[(f(x)-h(x))^2] = E[g(x)^2] = \frac{1}{2^n}\sum g(x)^2 = \langle g, g\rangle$

$= \sum_S \hat{g}(S)^2$ Parseval

$= \sum_{|S|\le d} \hat{g}(S)^2 + \sum_{|S|>d} \hat{g}(S)^2$

$\hat{g}(S)^2 \le \gamma^2$; $\le (\hat{d})\cdot\gamma^2$

$\le \varepsilon$ by F.C.

$\le \gamma + \varepsilon$ ∎

2nd "stage":

Thm 2 $f: \{\pm 1\}^n \to \{\pm 1\}$

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

then $\Pr[f(x) \neq \text{sign}(h(x))] \le E_{x\in U}[(f(x)-h(x))^2]$

Proof.

$$\mathop{E}_{x \in \mathcal{U}}\left[(f(x)-h(x))^2\right] = \frac{1}{2^n}\sum_x (f(x)-h(x))^2 \qquad \text{defn.}$$

$$\Pr\left[f(x) \neq \text{sign}(h(x))\right] = \frac{1}{2^n}\sum_x 1_{\{f(x)\neq \text{sign}(h(x))\}}$$

compare these term-by-term to get Thm.

Consider "$(f(x)-h(x))^2$" vs. "$1_{\{f(x)\neq \text{sign}(h(x))\}}$":

Case 1 if $f(x) = \text{sign}(h(x))$:

$1_{f(x)\neq \text{sign}(h(x))} = 0$

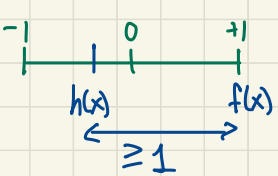
$(f(x)-h(x))^2 \geq 0$ ← bigger!

Case 2 if $f(x) \neq \text{sign}(h(x))$:

$1_{f(x)\neq \text{sign}(h(x))} = 1$

bigger! → $(f(x)-h(x))^2 \geq 1$

Why? e.g. if $f(x)=+1$ then in this case $h(x)<0$:



(other case is analogous)

So, $\forall x$

$(f(x)-h(x))^2 \geq 1_{f(x)\neq \text{sign}(h(x))}$ ∎

Correctness of learning algorithm

Thm if $\mathcal{C}$ has Fourier concentration $d = \alpha(\varepsilon, n)$

then there is a $q = O\left(\frac{n^d}{\varepsilon} \log \frac{n}{\delta}^d\right)$ sample uniform distribution learning algorithm for $\mathcal{C}$

ie. algorithm gets q samples of f with prob $\geq 1-\delta$

outputs h' s.t. $\Pr[f \neq h'] \leq 2\varepsilon$

Pf.

run low degree alg with $\gamma = \varepsilon$

thm 1 $\Rightarrow$ get h s.t. $E[(f-h)^2] \leq \varepsilon + \varepsilon = 2\varepsilon$

output $h' = \text{sign}(h)$

$\Uparrow$

thm 2 $\Rightarrow h'$ has error $\leq 2 \cdot \varepsilon$

Applications

1) Bounded depth decision trees

$$f(x) = \sum_{\ell \in \text{leaves of } T} f_\ell(x) \cdot \text{val}(\ell)$$

$f_\ell(x)$: fctn which depends on $\leq$ depth many vars

val(ℓ): const

$$\hat{f}(S) = \sum \text{val}(\ell)\, \hat{f}_\ell(S)$$ linearity

$\hat{f}_\ell(S)$: 0 for $|S| >$ depth

$$\Rightarrow \forall\, S \text{ s.t. } |S| > \text{depth}, \quad \hat{f}(S) = 0$$

so $O\left(\frac{n}{\varepsilon}^{\text{depth}} \log \frac{n}{\delta}^{\text{depth}}\right)$ suffices

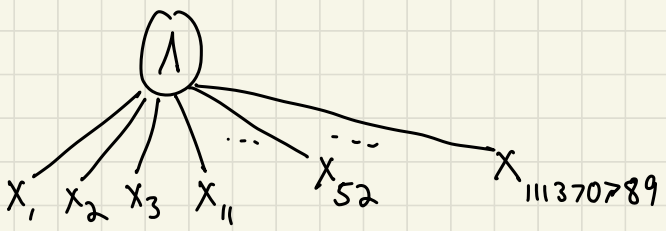
2) Constant depth ckts

def. "Boolean Ckt C" is DAG

gates: $\wedge, \vee, \neg$, $1, 0$, $x_1 \cdots x_n$

operations | consts | vars

how many inputs? const, poly, unbounded?



can we compute parity (xor) of n bits in const depth?

yes! can compute any fcn on n bits in const depth "Karnaugh maps"

parity in const depth, poly size?

no! [Furst Saxe Sipser] & Lemon Switching lemma

lemons ⇒ lemonade:

Thm [Hastad, Linial Mansour Nisan]

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

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

take $s = \text{poly}(n)$, $d = \text{const}$, $\alpha = O(\varepsilon)$ $\Rightarrow t = O\left(\log^d\left(\frac{n}{\varepsilon}\right)\right)$

yields $n^{O\left(\log^d\left(\frac{n}{\varepsilon}\right)\right)}$ sample algorithm

(can improve to $n^{O(\log \log n)}$ [Jackson])

(recall parity of s will have 1 large fourier coeff of degree $|s|$)

3) Learning halfspaces

def. $h(x) = \text{sign}(w \cdot x - \theta)$ is "halfspace function"

$$\text{sign}(y) = \begin{cases} +1 & \text{if } y \geq 0 \\ -1 & \text{o.w.} \end{cases}$$

Thm Let h be halfspace over $\{\pm 1\}^n$
then h has f.c. $\alpha(\varepsilon) = \frac{c}{\varepsilon^2}$

$$\left(\text{i.e. } \sum_{|S| \geq \frac{c}{\varepsilon^2}} \hat{h}(S)^2 \leq \varepsilon\right)$$

(will prove soon)

Corr low degree alg learns halfspaces under unif dist with $n^{O(1/\varepsilon^2)}$ unif. samples.

(actually $O(n^5)$ sample algorithms exist, but this approach will have "big win" soon)