Lecture 17

Fourier-based learning algorithms

- learning one Fourier coeff
- the low degree algorithm

Review from last time:

<u>def</u> 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 $\geq 1-\delta$
  $\underbrace{\text{error}(h) \text{ w.r.t. } f \text{ is} \leq \varepsilon}_{h \text{ is } \varepsilon\text{-close to } f}$

<u>Parameters of interest:</u>

- m # samples used by $\mathcal{A}$ "Sample complexity"
- ε accuracy parameter
- δ confidence parameter
- runtime? hope for poly(log(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 $\underbrace{\text{compact}}_{O(\log|\mathcal{C}|)}$ + efficient to evaluate

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:

<u>lemma</u> for any $S \subseteq [n]$, can approx $\hat{f}(S)$ to within additive γ

$$(\text{i.e. } |\text{output} - \hat{f}(S)| \le \gamma)$$

with prob $\ge 1-\delta$ in $O\left(\frac{1}{\gamma^2}\log\frac{1}{\delta}\right)$ samples.

no queries needed!

<u>Pf.</u> Chernoff + $\hat{f}(S) = 2\underbrace{\Pr_x[f(x) = \chi_S(x)]}_{\text{estimate this}} - 1$ ∎

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 <u>any</u> of them.
Need $\delta \ll \frac{1}{2^n}$, which needs $O(\frac{1}{\gamma^2} \cdot n)$ samples, but <u>exponential</u> 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!

Fourier Representations of Important Examples

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

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

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

AND over $\{0,1\}$ range

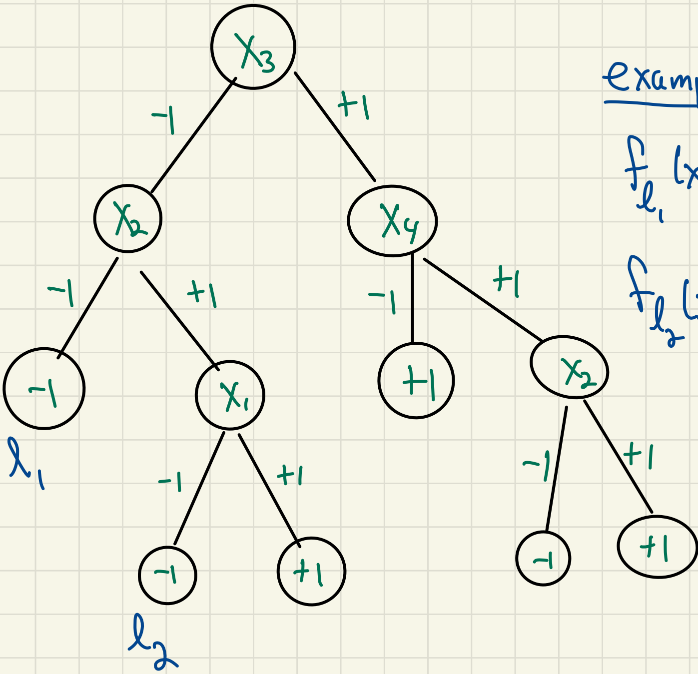
$$= \frac{(1-x_{i_1})}{2} \cdot \frac{(1-x_{i_2})}{2} \cdot \cdots \cdot \frac{(1-x_{i_k})}{2}$$

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

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

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:

$$f_\ell(x) = \prod_{\substack{i \in V_\ell \\ \text{vars visited on path to leaf } \ell}} \frac{(1 \overset{\text{left or right}}{\pm} x_i)}{2}$$

$$= \frac{1}{2^{|V_\ell|}} \sum_{S \subseteq V_\ell} (\overset{(-1)^{\#\text{ left turns taken in } S}}{\pm 1}) \chi_S = \begin{cases} 1 & \text{if } x \text{ takes path to } \ell \\ 0 & \text{o.w.} \end{cases}$$

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

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

Comment only coeffs corresponding to S s.t. $|S| \leq$ 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 $\displaystyle\sum_{\substack{S \subseteq [n] \\ \text{s.t.} \\ |S| > \alpha(\varepsilon, n)}} \hat{f}(S)^2 \leq \varepsilon \quad \forall\ 0 < \varepsilon < 1$

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 $\displaystyle\sum_{\substack{S \text{ s.t.} \\ |S| > k}} \hat{f}(S)^2 = 0$

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

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

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

$$\hat{f}(\varphi)^2 = (1 - 2\Pr(f(x) \neq \chi_\phi(x)))^2$$
$$= \left(1 - \frac{2}{2^{|T|}}\right)^2$$
$$> 1 - \varepsilon$$

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

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

Low degree algorithm

approximates fcfns with $d \equiv d(\varepsilon, n)$ Fourier concentration:

Given: d degree

τ accuracy

δ confidence

Algorithm:

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

  $C_S \leftarrow$ estimate of $\hat{f}(S)$ $\quad$ reuse samples

- let $h(x) \equiv \sum_{|S| \leq d} C_S \cdot \chi_S(x)$
- output $\text{sign}(h)$ as hypothesis

Why does this work?

Two stages:

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

1) Show that f has low F.C.

$$\Rightarrow \quad E_x[(f(x)-h(x))^2] \text{ small}$$

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

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 $E_x[(f(x)-h(x))^2 \leq \varepsilon + \tau$
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. $|s| \leq d$

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

<u>Pf of claim</u> (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 $\Pr\left[\,|C_S - \hat{f}(S)| > \gamma\right] < \frac{\delta}{n^d}$

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

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

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

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

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. $|S| > d$, $\hat{h}(S) = 0 \Rightarrow \hat{g}(S) = \hat{f}(S)$

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

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

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

So $$\mathbb{E}[(f(x)-h(x))^2] = \mathbb{E}[g(x)^2]$$

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

$$= \underbrace{\sum_{|S|\le d} \underbrace{\hat{g}(S)^2}_{\le \gamma^2}}_{\le \binom{n}{d}\cdot \gamma^2} + \underbrace{\sum_{|S|>d} \hat{g}(S)^2}_{\le \varepsilon \text{ by F.C.}}$$

$$\le \gamma + \varepsilon \qquad \blacksquare$$

2nd "stage":

<u>Thm 2</u> $f : \{\pm 1\}^n \to \{\pm 1\}$

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

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