6.842 Randomness and Computation April 6, 2022

Lectures on Learning Theory

Lecturer: Ronitt Rubinfield *Scribe: Yuchong Pan*

1 PAC Learning

The model of *learning from random, uniform examples* is as follows: Given the *example oracle* $\mathrm{Ex}(f)$ of a function f, pick m i.i.d. random variables $x_1, \ldots, x_m$ uniformly (or from some distribution $\mathcal{D}$, which might not be known to the learner in general), and call $\mathrm{Ex}(f)$ to obtain m random labeled examples $(x_1, f(x_1)), \ldots, (x_m, f(x_m))$; after seeing these examples, the learner outputs a hypothesis h of the function f.

Should we require $h = f$? This is probably too much to ask. However, we can at least require $\mathrm{dist}(h, f) := \mathbb{P}_{x\sim\mathcal{D}}(h(x) \neq f(x)) \leq \varepsilon$, where $\mathrm{dist}(h, f)$ is also called $\mathrm{error}_{\mathcal{D}}(h)$ with respect to f.

Definition 1. A *uniform distribution learning algorithm for a concept class* $\mathcal{C}$ is an algorithm $\mathcal{A}$ that, given $\varepsilon > 0$, $\delta > 0$ and access to $\mathrm{Ex}(f)$ for $f \in \mathcal{C}$, outputs a function h such that with probability at least $1 - \delta$, $\mathrm{error}(h)$ with respect to f is at most ε. This is called *probably approximately correct (PAC) learning.*

We are interested in the following parameters:

- m, the *sample complexity*;
- ε, the *accuracy* parameter;
- δ, the *confidence* parameter;
- the running time, which we hope to be $\mathrm{poly}(\log(\text{domain size}), 1/\varepsilon, 1/\delta)$;
- the *description* of h, which at least should be compact (i.e., $O(\log |\mathcal{C}|)$) and efficient to evaluate; it require $h \in \mathcal{C}$, then this is called *proper learning.*

Note that the uniform case is a special case of the PAC model. The more general PAC model is given $\mathrm{Ex}_{\mathcal{D}}(f)$ and bounds $\mathrm{error}_{\mathcal{D}}(h)$ with respect to f.

2 Learning Conjunctions

Let $\mathcal{C}$ be the class of conjunctions (i.e., 1-term DNF) over $\{0,1\}^n$. We cannot hope for 0-error from a sub-exponential number of random samples; to see this, note that it is hard to distinguish $f(x) = x_1 \cdots x_n$ and $f(x) = \mathsf{F}$. Algorithm 1 gives a polynomial time sampling algorithm for conjunction learning, where "**?**" indicates a parameter to be determined.

For x_i in the conjunction, it must be set in the same way in each positive example, so $i \in V$. For x_i not in the conjunction,

$$\mathbb{P}[i \in V] = \mathbb{P}\left[x_i \text{ is set is the same way in each of the } k \text{ positive examples}\right] = \frac{1}{2^{k-1}}.$$

By the union bound,

$$\mathbb{P}\left[\text{any } x_i \text{ not in the conjunction survives}\right] \leq \frac{n}{2^{k-1}} \leq \delta,$$

draw poly$(1/\varepsilon)$ samples
estimate $\mathbb{P}[f(x) = 1]$ to additive error at most $\pm\varepsilon/4$ and confidence at least $1 - \delta/2$
if *estimate is less than* $\varepsilon/2$ **then**
    return $h(x) = 0$
(estimate is at least $\varepsilon/2$; see a new positive example every $O(1/\varepsilon)$ samples)
collect ? more positive examples
$V \leftarrow$ set of indices of variables that are set in the same way in each positive example
return $h(x) = \bigwedge_{i \in V} x_i^{b_i}$, *where each* b_i *indicates if* x_i *is complemented or not*

Algorithm 1: A polynomial time sampling algorithm for conjunction learning.

if we pick $k = \log(n/\delta)$. Therefore, if we need $\Omega(\log(n/\delta))$ positive examples, or $\Omega((1/\varepsilon)\log(n/\delta))$ total examples to rule out every x_i not in the conjunction.

3 Occam's Razor

In a high level, *Occam's Razor* claims the following:

- If we ignore the running time, then learning is easy (with a polynomial number of samples).
- The shortest explanation is the best.

To see the first claim, we consider the brute-force algorithm in Algorithm 2.

draw $M = (1/\varepsilon)(\ln|\mathcal{C}| + \ln|1/\delta|)$
search over all $h \in \mathcal{C}$ until find one consistent with the samples
return h

Algorithm 2: A brute-force learning algorithm that demonstrates Occam's Razor.

We say that a function h is *bad* if error(h) with respect to f is at least ε. For a bad function h,

$$\mathbb{P}[h \text{ is consistent with the samples}] \leq (1-\varepsilon)^M.$$

By the union bound,

$$\mathbb{P}[\text{any bad function } h \text{ is consistent with the samples}] \leq |\mathcal{C}|(1-\varepsilon)^M = |\mathcal{C}|(1-\varepsilon)^{\frac{1}{\varepsilon}\left(\ln|\mathcal{C}| + \ln\left|\frac{1}{\delta}\right|\right)} = \delta.$$

Hence, it is unlikely to output a bad hypothesis h. For example, for conjunction learning, this analysis requires $O((1/\varepsilon)(n + 1/\delta))$ samples, where Algorithm 1 has a better sample complexity. On the other hand, if we have a *good* hypothesis h,

(i) we can *predict* values of f on new random inputs according to distribution $\mathcal{D}$, since

$$\mathbb{P}_{x \sim \mathcal{D}}[f(x) = h(x)] \geq 1 - \delta;$$

(ii) we can *compress* the description of samples $(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m))$ from the naïve description which takes $m(\log|D| + \log|R|)$ bits, where D and R are the domain and the range of f, respectively, to the form "$x_1, \ldots, x_m$ plus the description of h" which requires $m \log|D| + \log|\mathcal{C}|$ bits only.

4 Learning via Fourier Representations

In this section, we study learning algorithms that are based on estimating the Fourier representation of a function f.

4.1 Approximating One Fourier Coefficient

Lemma 2. *For any $S \subset [n]$, one can approximate $\hat{f}(S)$ to within additive error γ (i.e., $|\text{output} - \hat{f}(S)| \leq \gamma$) with probability at least $1 - \delta$ in $O(1/\gamma^2 \log 1/\delta)$ samples.*

Proof. Recall that $\hat{f}(S) = 2\,\mathbb{P}_{x\in\{0,1\}^n}[f(x) \neq \chi_S(x)] - 1$. Hence, we can approximate $\hat{f}(S)$ by estimating $\mathbb{P}_{x\in\{0,1\}^n}[f(x) \neq \chi_S(x)]$ and applying the Chernoff bound. □

We examine the Fourier representations of some important examples:

(i) Let $T = \{i_1, \ldots, i_k\} \subset [n]$ be such that $|T| = k$. Let $\overline{\mathsf{AND}} : \{\pm 1\}^n \to \{\pm 1\}$ be defined by

$$\overline{\mathsf{AND}}(x) = \begin{cases} 1, & \text{if } x_i = -1 \text{ for all } i \in T, \\ -1, & \text{otherwise.} \end{cases}$$

Let $f : \{\pm 1\}^n \to \{0, 1\}$ be defined by

$$\begin{aligned} f(x) &= \begin{cases} 1, & \text{if } x_i = -1 \text{ for all } i \in T, \\ 0, & \text{otherwise,} \end{cases} \\ &= \frac{1 - x_{i_1}}{2} \cdot \frac{1 - x_{i_2}}{2} \cdots \frac{1 - x_{i_k}}{2} \\ &= \frac{1}{2^k} \sum_{S \subset T} (-1)^{|S|} \chi_S(x). \end{aligned}$$

Note that all Fourier coefficients $\hat{f}(S)$ for $S \not\subset T$ are 0. Then

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

(ii) ***Decision trees.*** Consider a decision tree T, e.g., Figure 1. For each leaf ℓ of T, let V_ℓ denote the set of indices of variables visited on the path from the root to leaf ℓ, and let $f_\ell : \{\pm 1\}^n \to \{0, 1\}$ be defined by

$$\begin{aligned} f_\ell(x) &= \prod_{i \in V_\ell} \frac{1 \pm x_i}{2} && (\text{“}\pm\text{” denotes a left turn or a right turn}) \\ &= \frac{1}{2^{|V_\ell|}} \sum_{S \subset V_\ell} (-1)^{\#\text{ left turns taken in } S} \chi_S. \end{aligned}$$

Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be defined by

$$f(x) = \sum_{\ell\text{: leaf of } T} f_\ell(x)\,\mathrm{val}(\ell).$$

Note that for each $x \in \{\pm 1\}^n$, exactly one of the values $f_\ell(x)$ is 1 for leaves ℓ of T, and all others are 0.

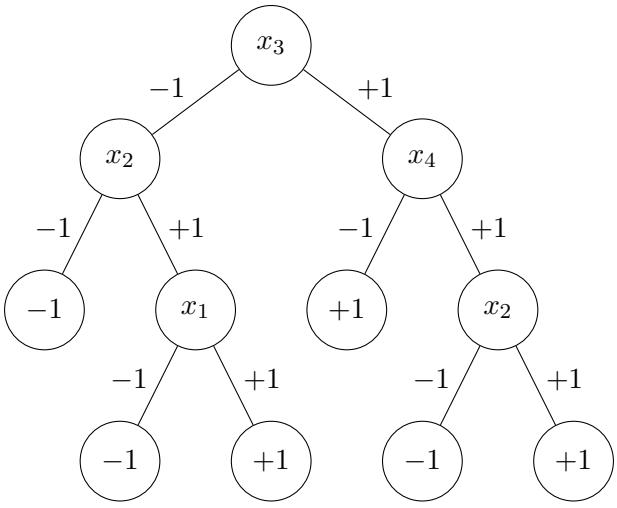

Figure 1: A decision tree.

4.2 Fourier Concentration

Definition 3. For all $\varepsilon \in (0,1)$ and $n \in \mathbb{N}$, we say that a function $f : \{\pm 1\}^n \to \mathbb{R}$ has $\alpha(\varepsilon, n)$-*Fourier concentration (f.c.)* if

$$\sum_{\substack{S \subset [n] \\ |S| > \alpha(\varepsilon,n)}} \hat{f}(S)^2 \leq \varepsilon.$$

For a Boolean function $f : \{\pm 1\}^n \to \{\pm 1\}$, Parseval's identity gives $\sum_{S \subset [n]} \hat{f}(S)^2 = 1$, so having $\alpha(\varepsilon, n)$-Fourier concentration implies that

$$\sum_{\substack{S \subset [n] \\ |S| \leq \alpha(\varepsilon,n)}} \geq 1 - \varepsilon.$$

Following are two examples of functions that have $\alpha(\varepsilon, n)$-Fourier concentration.

(i) Any function $f : \{\pm 1\}^n \to \mathbb{R}$ that depends on at most k variables has

$$\sum_{\substack{S \subset [n] \\ |S| > k}} \hat{f}(S)^2 = 0.$$

(ii) Let f be the $\overline{\mathsf{AND}}$ function on $T \subset [n]$. We claim that f has $\log(4/\varepsilon)$-Fourier concentration. Note $\hat{f}(S) = 0$ for all $S \subset [n]$ with $|S| > |T|$. If $|T| \leq \log(4/\varepsilon)$, then we are done by definition. If $|T| > \log(4/\varepsilon)$, then

$$\hat{f}(\emptyset)^2 = (1 - 2\,\mathbb{P}\left[f(x) \neq \chi_\emptyset(x)\right])^2 = \left(1 - 2 \cdot \frac{1}{2^{|T|}}\right)^2 > 1 - \varepsilon.$$

Therefore, f has 0-Fourier concentration.

take $m = O((n^d/\tau)\log(n^d/\delta))$ samples
foreach $S \subset [n]$ *such that* $|S| \le d$ **do**
    $C_S \leftarrow$ estimate of $\hat{f}(S)$
let $h : \{\pm 1\}^n \to \mathbb{R}$ be defined by $h(x) = \sum_{S\subset[n]:|S|\le d} C_S \chi_S(x)$
return $\operatorname{sign} \circ h$ *as hypothesis*

Algorithm 3: The low degree algorithm given degree d, accuracy τ and confidence δ.

4.3 The Low Degree Algorithm

The *low degree algorithm*, given in Algorithm 3, approximates a Boolean function with d-Fourier concentration, where $d = \alpha(\varepsilon, n)$.

Proposition 4. *If f has d-Fourier concentration with $d = \alpha(\varepsilon, n)$, then $\mathbb{E}_{x\in\{0,1\}^n}[(f(x)-h(x))^2] \le \varepsilon + \tau$ with probability at least $1-\delta$.*

Proof. First, we claim that each low degree Fourier Coefficient is well approximated, i.e., with probability at least $1-\delta$, we have $|C_S - \hat{f}(S)| \le \gamma$ for all $S \subset [n]$ with $|S| \le d$, where $\gamma = \sqrt{\tau/n^d}$. This can be proved using the Chernoff bound and the union bound.

Second, we show that if all low degree Fourier coefficients are well approximated, then h has a low ℓ_2-error. Suppose $|C_S - \hat{f}(S)| \le \gamma$ for all $S \subset [n]$ such that $|S| \le d$. Let $g : \{\pm 1\}^n \to \mathbb{R}$ be defined by

$$g(x) = f(x) - h(x).$$

By the linearity of the Fourier transform, for all $S \subset [n]$,

$$\hat{g}(S) = \hat{f}(S) - \hat{h}(S).$$

For all $S \subset [n]$ with $|S| > d$, we have $\hat{h}(S) = 0$, so $\hat{g}(S) = \hat{f}(S)$. For all $S \subset [n]$ with $|S| \le d$, we have $|\hat{g}(S)| \le \gamma$, so $\hat{g}(S)^2 \le \gamma^2$. Therefore,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\in\{\pm1\}^n}\left[(f(x)-h(x))^2\right] &= \mathbb{E}\left[g(x)^2\right] = \sum_{S\subset[n]} \hat{g}(S)^2 && \text{(Parseval's identity)} \\
&= \sum_{\substack{S\subset[n]\\|S|\le d}} \hat{g}(S)^2 + \sum_{\substack{S\subset[n]\\|S|>d}} \hat{g}(S)^2 \\
&\le \sum_{\substack{S\subset[n]\\|S|\le d}} \gamma^2 + \sum_{\substack{S\subset[n]\\|S|>d}} \hat{f}(S)^2 \\
&\le \binom{n}{d}\cdot\gamma^2 + \varepsilon && \text{(by Fourier concentration)} \\
&\le \tau + \varepsilon.
\end{aligned}$$

This completes the proof. □

Proposition 5. *Let $f : \{\pm1\}^n \to \{\pm1\}$. Let $h : \{\pm1\}^n \to \mathbb{R}$. Then*

$$\mathop{\mathbb{P}}_{x\in\{\pm1\}^n}[f(x) \ne \operatorname{sign}(h(x))] \le \mathop{\mathbb{E}}_{x\in\{\pm1\}^n}\left[(f(x)-h(x))^2\right]$$