CS 7545: Machine Learning Theory **Fall 2021**

Lecture 14: Statistical Query Model

Instructor: Santosh Vempala *Lecture date: 11/08-10/2021*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

We have seen PAC learning algorithms that work when there is no noise in the input data. Suppose there exists h which correctly classifies 99% of D and we want a function $\tilde{h}$ with accuracy 51%? (You can get 50% accuracy by guessing randomly.)

Random Classification Noise Model In this model, the learner is given labelled examples $(x, \ell(x))$ where

$$\ell(x) = \begin{cases} f(x) & \text{w.p. } 1-\eta \\ 1-f(x) & \text{w.p. } \eta. \end{cases}$$

Can we still learn the underlying concept?

For example, consider the class of Boolean OR functions. Without noise, a PAC learning algorithm is to remove all variables that are 1 when $\ell(x) = 0$ and remove the disjunction of all remaining variables.

With η fraction of noise, we can do the following: let $p_i = \Pr(f(x) = 0 \text{ and } x_i = 1)$.

- Include all x_i for which $p_i = 0$
- Do not include any x_i for which $p_i \geq \epsilon/n$

For this algorithm,

$$\Pr(\text{correct}) \geq 1 - \frac{\epsilon}{n} \cdot n = 1 - \epsilon.$$

Can you estimate p_i to within $\pm\epsilon/n$ error? Note that

$$\Pr(f(x) = 0 \text{ and } x_i = 1) = \underbrace{\Pr(f(x) = 0 | x_i = 1)}_{q_i} \cdot \underbrace{\Pr(x_i = 1)}_{\text{independent of noise}}.$$

In the random noise model,

$$\begin{aligned} p_i = \Pr_{\eta}(\ell(x) = 0 | x_i = 1) &= (1-\eta) \cdot \Pr(f(x) = 0 | x_i = 1) + \eta \cdot \Pr(f(x) = 1 | x_i = 1) \\ &= \eta q_i + \eta(1 - q_i) = \eta + q_i(1 - 2\eta). \end{aligned} \tag{14.1}$$

So, $q_i = \frac{p_i - \eta}{1-2\eta}$ and it suffices to estimate p_i to $\pm\frac{\epsilon}{n}(1-2\eta)$ error.

14.1 Statistical Query Model

In this model, the algorithm can ask an oracle for expectation of any bounded function of $(x, \ell(x))$ upto an additive error τ. So, the algorithm can query the oracle with

$$h : X \times \{0,1\} \to [0,1],\ \tau > 0$$

and the SQ oracle responds with $\mathbb{E}_D(h) \pm \tau$. Here h should be polytime computable and $\tau \geq 1/poly$.

Example: The algorithm can find $\mathbb{E}(x_i | \ell(x) = 0)$ by using $h(x, \ell(x)) = x_i$ if $\ell(x) = 1$ and 0 otherwise. Many known learning algorithms can be implemented with SQ. For example, in gradient descent, the loss function is $\mathcal{L}(w) = \mathbb{E}_{x,y}(\ell(x, y, w))$ and the gradient is $\nabla_w \mathcal{L}(w) = \mathbb{E}_{x,y}(\nabla_w \ell(x, y, w))$.

Theorem 14.1. *PAC learning in Statistical Query model ⇒ PAC learning with noise.*

Proof. For a function $h : X \times \{0,1\} \to \{0,1\}$, let

$$\begin{aligned}\text{CLEAN} &= \{x : h(x,0) = h(x,1)\}\\ \text{NOISY} &= \{x : h(x,0) \neq h(x,1)\}.\end{aligned}$$

Then $\mathbb{E}(h(x,f(x))) = \Pr(h(x,f(x)) = 1)$ and

$$\begin{aligned}\Pr(h(x,f(x))=1) &= \Pr(h(x,f(x))=1 \;\&\; x \in \text{CLEAN}) + \Pr(h(x,f(x))=1 \;\&\; x \in \text{NOISY})\\ &= \underbrace{\Pr(h(x,f(x))=1 \;\&\; x \in \text{CLEAN})}_{\text{unaffected by noise}} + \underbrace{\Pr(h(x,f(x))=1 | x \in \text{NOISY})}_{p} \cdot \Pr(x \in \text{NOISY})\\ &= \Pr(h(x,f(x))=1 \;\&\; x \in \text{CLEAN}) + p \cdot \underbrace{(1 - \Pr(x \in \text{CLEAN}))}_{\text{unaffected by noise}}\end{aligned}$$

In the random noise model, the algorithm can estimate $q = \Pr_\eta(h(x,\ell(x)) = 1 | x \in \text{NOISY})$ by querying the SQ oracle. Similar to (14.1), $q = (1-\eta)p + \eta(1-p) = \eta + p(1-\eta)$. To estimate $p \pm \tau$, it suffices to estimate $q \pm \tau(1-2\eta)$. □

Some functions are hard to learn in the SQ model. Suppose the hypothesis class is the set of parity functions on n variables. The following theorem shows that we need $2^{\Omega(n)}$ SQ calls to learn a function from this class.

Theorem 14.2. *Weakly learning PARITY requires $2^n\tau^2$ queries to STAT(τ), where $\tau \geq 1/poly(n)$.*

Proof. Any SQ $h(x,\ell(x))$ is a function from $\{-1,1\} \times \{-1,1\}^n \to [-1,1]$. For any $S \subseteq [n]$, let $h_S(x,f(x)) = \chi_S(x)f(x)$ and $\chi_S(x,f(x)) = \chi_S(x)$. Then, $\{h_S(x,f(x))\} \cup \{\chi_S(x,f(x))\}$ is a set of 2^{n+1} functions and

- $\langle \chi_S, \chi_T \rangle_D = \begin{cases} 0 & \text{if } S \neq T \\ 1 & \text{if } S = T \end{cases}$
- $\langle h_S, h_T \rangle_D = \langle \chi_S, \chi_T \rangle_D = \begin{cases} 0 & \text{if } S \neq T \\ 1 & \text{if } S = T \end{cases}$
- $\langle \chi_S, h_T \rangle_D = \langle \chi_S, f(x)\chi_T \rangle_D = 0$ for any S, T

Therefore $\{h_S(x,f(x))\}, \{\chi_S(x,f(x))\}$ form an orthonormal basis for $F : \{-1,1\}^n \times \{-1,1\} \to \mathbb{R}$. So any function $g : \{-1,1\}^n \times \{-1,1\} \to [-1,1]$ can be written as

$$g(x,f(x)) = \sum_S \alpha_S h_S(x,f(x)) + \sum_S \hat{g}_S \chi_S(x,f(x)).$$

Since

$$\mathbb{E}(h_S(x,f(x))) = \begin{cases} 1 & \text{if } S = S^* \\ 0 & \text{otherwise,} \end{cases}$$

we get

$$\mathbb{E}(g(x,f(x))) = \sum_S \alpha_S \mathbb{E}(h_S(x,f(x))) + g_0 = \sum_S \alpha_S \mathbb{E}(f(x)\chi_S) + g_0 = \alpha_{S^*} + g_0.$$

where g_0 is a constant independent of $f(x)$ and S^*.

The SQ oracle can output g_0 for every query as long as $\alpha_S^* \leq \tau$. Since $\|h\|_D^2 = 1$, the number of S with $\alpha_S \geq \tau$ is at most $1/\tau^2$. Therefore, for every query, the oracle eliminates at most $1/\tau^2$ subsets from the set of possible candidates for the parity function. So, the algorithm needs to make at least $2^n/(1/\tau^2) = 2^n\tau^2$ queries to learn anything about the parity function. □

Similar lower bound holds when the label is noisy. More generally,

$$SQ - dim(\mathcal{H}, \gamma) = \max\{d : \exists d \text{ concepts } h_1, h_2, \ldots, h_d \in \mathcal{H} \text{ s.t. } \|h_i\|^2 \leq 1 \text{ and } \langle h_i, h_j \rangle_D \leq \gamma\}.$$

Theorem 14.3. *To learn $\mathcal{H}$ with $SQ - dim(\mathcal{H}) = d$ in the SQ model, we need $\Omega(d\gamma)$ queries to $SQ(\sqrt{\gamma})$.*