# Lecture 12: Learning with RCN and Statistical Learning

*Instructor: Vatsal Sharan* *Scribe: Neel Patel*

In previous class, we discussed the definition of efficient PAC learning with Random Classification Noise and Statistical Query learning model. We can describe noisy oracle as follows:

**Oracle:** $\text{EX}^\eta$

- Draws $X \sim \mathcal{D}$ from $\mathcal{X}$
- With probability $(1-\eta)$, return $(x, c(x))$ otherwise flips the label and return $(x, 1-c(x))$

We can describe statistical query oracle as follows:

**Oracle:** Query $(\phi, \tau)$

- $\phi : \mathcal{X} \times \{0,1\}$ (Query function) and $\tau \in (0,1)$ (query tolerance)
- Let $p_\phi = \Pr_{X\sim D}[\phi(X, c(X)) = 1]$
- Statistical oracle returns $\hat{p}_\phi$ such that $\hat{p}_\phi \in (p_\phi - \tau, p_\phi + \tau)$

## 1 SQ Learning $\implies$ PAC Learning in presence of RCN

In this section, we will show that if a concept class is efficiently SQ learnable then it is also PAC learnable. We formalize our result in the following theorem.

**Theorem 1.** *If concept class $\mathcal{C}$ is efficiently SQ learnable then $\mathcal{C}$ is PAC learnable in the presence of random classification noise.*

*Proof.* In order to prove the theorem, we need to show that given access to $\text{EX}^\eta$, we can simulate $\text{STAT}(c, D)$ oracle with bounded (low) failure probability. The key idea is that given SQ oracle, we can divide domain $\mathcal{X}$ into two disjoint parts:

1. $\mathcal{X}_1$ : all $x \in \mathcal{X}$ such that $\phi(x, 0) \neq \phi(x, 1)$, data region where output of SQ is dependent on label.
2. $\mathcal{X}_2$ : all $x \in \mathcal{X}$ such that $\phi(x, 0) = \phi(x, 1)$, data region where output of SQ is not dependent on label.

We further define conditional data distributions on these defined regions. Let $D_1$ be the conditional distribution of $x$ restricted to $\mathcal{X}_1$ with $p_1 = \Pr_{x\sim D}[x \in \mathcal{X}_1]$ and $D_2$ be the conditional distribution of

$x$ restricted to $\mathcal{X}_2$ with $p_2 = \Pr_{x \sim D}[x \in \mathcal{X}_2]$. We first decompose $p_\phi$:

$$\begin{aligned} p_\phi &= \mathbb{E}_{EX(c,D)}[\phi(x, c(x))] \\ &= \Pr_{EX(c,D)}[x \in \mathcal{X}_1] \cdot \Pr_{EX(c,D)}[\phi = 1 | x \in \mathcal{X}_1] + \Pr_{EX(c,D)}[x \in \mathcal{X}_2] \cdot \Pr_{EX(c,D)}[\phi = 1 | x \in \mathcal{X}_2] \\ &= p_1 \cdot \Pr_{EX(c,D)}[\phi = 1 | x \in \mathcal{X}_1] + \Pr_{EX(c,D)}[\phi = 1 \wedge x \in \mathcal{X}_2] \end{aligned}$$

Now, in order to prove the theorem, we need to show that we can approximate all the terms in the above decomposition with a small approximation error using a noisy oracle. Note that $\phi$ does not depend on the label of $x$ in the region $\mathcal{X}_2$. Hence, we can approximate $\Pr_{EX(c,D)}[\phi = 1 \wedge x \in \mathcal{X}_2]$ using noisy oracle in poly many noisy queries because the event $\phi = 1$ is independent of the labels.

More formally, we can use rejection sampling. We sample $(x \sim \mathcal{D}, c(x))$ using noisy oracle, if $x \in \mathcal{X}_2$ and $\phi(x) = 1$ then we accept the sample, and otherwise, we reject the sample. We can compute fraction of accepted sample as an estimate of $\Pr[\phi = 1 \wedge x \in \mathcal{X}_2]$. Using concentration bounds, we can show that our estimate has error $O(\tau)$ with $O\left(\frac{1}{\tau^2}\log(1/\delta')\right)$ noisy queries.

Now, we can similarly compute $p_1$ using rejection sampling and obtain similar error rate: sample $(x \sim \mathcal{D}, c(x))$ using noisy oracle, if $x \in \mathcal{X}_1$ then we accept the sample, and otherwise, we reject the sample. We can compute fraction of accepted sample as an estimate of $p_1$. Using concentration bounds, we can show that our estimate has error $O(\tau)$ with $O\left(\frac{1}{\tau^2}\log(1/\delta')\right)$ noisy queries.

Now in order to approximate $\Pr_{EX(c,D)}[\phi = 1 | x \in \mathcal{X}_1] = \Pr_{EX(c,D_1)}[\phi = 1]$, for the sake of simplicity, we first assume that $\eta$ is already known. We can decompose the probability $\Pr_{EX^\eta(c,D_1)}[\phi = 1]$ as follows:

$$\begin{aligned} \Pr_{EX^\eta(c,D_1)}[\phi = 1] &= (1-\eta)\Pr_{EX(c,D_1)}[\phi = 1] + \eta \Pr_{EX(c,D_1)}[\phi = 0] \\ &= (1-\eta)\Pr_{EX(c,D_1)}[\phi = 1] + \eta(1 - \Pr_{EX(c,D_1)}[\phi = 1]) \\ &= \eta + (1-2\eta)\Pr_{EX(c,D_1)}[\phi = 1] \end{aligned}$$

Therefore,

$$\Pr_{EX(c,D_1)}[\phi = 1] = \frac{\Pr_{EX^\eta(c,D_1)}[\phi = 1] - \eta}{1 - 2\eta}$$

□

The above equation establish relation between $\Pr_{EX(c,D_1)}[\phi = 1]$ and $\Pr_{EX^\eta(c,D_1)}[\phi = 1]$ which allows us to approximate $\Pr_{EX^\eta(c,D_1)}[\phi = 1]$ using noisy oracle. Note that we in order to bound error by $O(\tau)$in the estimate of $\Pr_{EX(c,D_1)}[\phi = 1]$, we need to bound error in $\Pr_{EX^\eta(c,D_1)}[\phi = 1]$ by $\tau(1-2\eta)+\eta$. We can again use rejection sampling to estimate $\Pr_{EX^\eta(c,D_1)}[\phi = 1]$ similar to earlier cases and using concentration bounds, we can show that the estimation error of $\Pr_{EX(c,D_1)}[\phi = 1]$ is bounded by $O(\tau)$ with $O\left(\frac{1}{\tau^2(1-2\eta)^2}\log(1/\delta')\right)$ noisy queries.

We can repeat this for all SQ queries made by the SQ learning algorithm. We can set $\delta' = \delta/m$ and by union bound, we can bound the failure probability by $\delta$. However, now we have to get rid of our earlier assumption that we know $\eta$.

Suppose, we know that $\eta_0 \geq \eta$ (we can always assume that $\eta \leq 1/2 - \alpha$). Now, consider a small $\Delta > 0$, and construct $\Delta$-net for all possible values of $\eta$, i.e.

$$\Gamma = \left\{ i \cdot \Delta : o \leq i \leq \frac{\eta_0}{4} \right\}$$

Try all values of $\hat{\eta} \in \Gamma$. We know that at least one of the values of $\hat{\eta}$ will have a small error say $\epsilon/100$.

Note that we only need to estimate true error $\Pr_{EX(c,D)}[h(x) \neq c(x)]$. Let $h_i$ be the hypotheses produced in $i$-th iteration while passing through $\Gamma$ set: let $\gamma_i = \Pr_{EX^\eta(c,D)}[h_i(x) \neq c(x)]$

$$\gamma_i = (1-\eta) \Pr_{EX(c,D)}[h(x) \neq c(x)] + \eta(1 - \Pr_{EX(c,D)}[h(x) \neq c(x))$$

Therefore,

$$\Pr_{EX(c,D)}[h(x) \neq c(x)] = \frac{\gamma_i}{1-2\eta} - \frac{\eta}{1-2\eta}.$$

Hence, we can estimate $\gamma_i$ accurately for all $i$, and we choose best $h_i$ then we get our required error bound. By some cumbersome algebra, we can show that error is bounded by $O(\tau)$ in $O\left(\frac{1}{\Delta\tau^2(1-2\eta)^2}\log(m/\delta)\right)$ noisy queries. Concluding the proof.

## 2 Statistical Query Learnability and SQ Dimension

In this section, we characterize the learnability using SQ algorithms. We first define uncorrelated functions:

**Definition 2.** *Two functions $f, g$ defined on the same domain are uncorrelated if $\Pr_{x\sim D}[f(x) = g(x)]$.*

Now, we are ready to define *SQ* dimension.

**Definition 3.** *The SQ-dimension of a class $\mathcal{C}$ wrt. a distribution $D$ over $\mathcal{X}$ is the size of the largest subset $\mathcal{C}' \subset \mathcal{C}$ such that for all $f, g \in \mathcal{C}'$*

$$|\Pr_{x\sim D}[f(x) = g(x)] - 1/2| < 1/|C'|$$

**Definition 4** (Weak Learning)**.** *An algorithm $A$ is a weak learner with advantage $\gamma$ for class $\mathcal{C}$ if: for any dist. $D$ and any target $c \in \mathcal{C}$, given access to $EX(c, D)$, w.p. $(1-\delta)$, produces a hypotheses with error$(h; c, D) \leq \gamma$.*

In the next lecture, we will show that weak learning implies strong PAC learning. Now, we are ready to characterize SQ learnability for the SQ dimension.

**Theorem 5.** *If SQ-DIM$_D$ = poly$(d)$, then you can efficiently "weak learn" $\mathcal{C}$ over $D$ (get error rate $\leq 1/2 - 1/\text{poly}(d)$) using SQ-learning algorithm.*

*Proof.* Let $s =$SQ-DIM$_D(c)$, let $\mathcal{H} \subseteq \mathcal{C}$ be maximal subset such that $\forall h_i, h_j \in \mathcal{H}$, we have

$$\left|\Pr_D[h_i(x) = h_j(x)] - \frac{1}{2}\right| < \frac{1}{1+s}.$$

Therefore, $|\mathcal{H}| \leq s$. We try every $h_i \in \mathcal{H}$ and use SQ-oracle to estimate its error.

**Claim:** At least one $h_i$ or (complement or negation of $h_i$) must be a weak leaner.

Now, if target $c$ satisfied:

$$\left|\Pr_D[h_i(x) = c(x)] - \frac{1}{2}\right| < \frac{1}{s+1} \; \forall \; h_i \in \mathcal{H}$$

then we can include $c$ in the set $c$ in the set $\mathcal{H}$ which is a contradiction! Hence, there exists one weak learner in $\mathcal{H}$. □

**Theorem 6.** *If SQ-DIM$_D$ > poly$(d)$ then you cannot efficiently learn $\mathcal{C}$ over $D$ by SQ-algorithms.*

We will prove this theorem in the next class, however, we can use this theorem to show PARITY functions are not efficiently SQ learnable.

**Proposition 7.** *PARITIES are not efficiently SQ-Learnable.*

*Proof.* We can show that any two parity functions are uncorrelated for uniform distribution over $\{0,1\}^d$. Consider any two distinct PARITY functions $C_{w_1}(x)$ and $C_{w_2}(x)$: $C(x) = C_{w_1}(x) - C_{w_2}(x)$ (in modulo addition) is also a parity. Now, when each $x_i$ is 1 with probability 1/2, independently, $\Pr_{x\sim U}[C(x) = 0] = \Pr_{x\sim U}[C(x) = 1] = 1/2$. This implies that $\Pr_{x\sim U}[C(x) = 0] = 1/2$. Hence, PARITY functions are not efficiently SQ-learnable.

□

We can show that PARITIES are efficiently PAC learnable. This shows that there exists a concept class that is efficiently PAC learnable but not SQ learnable. Hence, efficient PAC learning does not imply efficient SQ learnability.

**Proposition 8.** *PARITIES are efficiently PAC learnable.*

*Proof.* First, note that since $|\mathcal{C}| = 2^d$, an ERM algorithm will get error $\epsilon$ woth $O\left(\frac{d \log 1/\delta}{\epsilon}\right)$. Now we show that ERM can be implemented in polynomial time. Given examples from any distribution $D$, $\{(a_1, b_1), \ldots, (a_n, b_n)\}$. It is easy to observe that ERM can be obtained by solving system of linear equation over field $\mathbb{F}_2$. □