# Notes 19: Lower bound for Statistical Query model

We saw that if $\mathcal{C}$ is efficiently learnable from SQ's, then $\mathcal{C}$ is efficiently PAC-learnable (even with RCN)
Question: If $\mathcal{C}$ is efficiently PAC-learnable, must $\mathcal{C}$ be efficiently learnable from SQ's?
Answer: No, a counterexample is $\mathcal{C} = \{\text{parity functions}\}$ over $X = \{0,1\}^n$
Recall a parity function $c(x) = \oplus_{i \in S} x_i$ for some $S \subseteq \{1, \dots, n\}$
e.g. $c(x) = x_1 \oplus x_2 \oplus x_4$ outputs the parity of the 1st, 2nd, 4th bits when $S = \{1,2,4\}$
$\mathcal{C} = \{\text{parity functions}\}$ is efficiently PAC-learnable using Gaussian elimination over $\mathbb{F}_2$

---

However, $\mathcal{C}$ is not efficiently learnable from SQ's
Fix $\mathcal{D}$ = uniform distribution over $X = \{+1,-1\}^n$ (note: switched from $\{0,1\}$ to $\{+1,-1\}$)
$\mathbb{E}_{\mathcal{D}}[f(x)g(x)]$ defines an **inner product** $\langle f, g\rangle$ between $f$ and $g$, where $f, g : X \to \mathbb{R}$

An orthogonal basis is $\{\mathbb{1}_x \mid x \in \{+1,-1\}^n\}$
i.e. $\mathbb{1}_x(z) = 1$ if $x = z$ and $\mathbb{1}_x(z) = 0$ if $x \neq z$
Orthogonal because $\langle \mathbb{1}_x, \mathbb{1}_y\rangle = \mathbb{E}_{z\in X}[\mathbb{1}_x(z)\mathbb{1}_y(z)] = 0$ whenever $x \neq y$

---

Better (orthonormal) basis: **Fourier basis** of parity functions $\{c_S \mid S \subseteq [n] = \{1, \dots, n\}\}$
Here $c_S : X \to \{+1,-1\}$ is given by $c_S(x) = \prod_{i\in S} x_i$
e.g. $c_{\{1,2,4\}}(z) = z_1 z_2 z_4$ and $c_{\{1,2,4\}}(+1, -1, -1, -1, +1) = (+1)(-1)(-1) = 1$
$c_\emptyset$ is the constant 1 function
We now show this basis is orthonormal, i.e. $\langle c_S, c_S\rangle = 1$ and $\langle c_S, c_T\rangle = 0$ for any $S \neq T$

$$\text{Expand} \quad \langle c_S, c_T\rangle = \mathop{\mathbb{E}}_{z\in\{+1,-1\}^n}[c_S(z)c_T(z)] = \mathop{\mathbb{E}}_{z\in\{+1,-1\}^n}\left[\textstyle\prod_{i\in S} z_i \prod_{i\in T} z_i\right]$$

Rewrite the factors inside the expectation as $\prod_{i\in S} z_i \prod_{i\in T} z_i = \prod_{i \in S \triangle T} z_i$ (because $z_i \in \{+1,-1\}$)
e.g. $S = \{1,2,3\}, T = \{3,4\}$, $(z_1z_2z_3)(z_3z_4) = z_1z_2z_3^2z_4 = z_1z_2z_4$ when $z \in \{+1,-1\}^n$
The expectation becomes

$$\mathop{\mathbb{E}}_{z\in\{+1,-1\}^n}\left[\textstyle\prod_{i\in S\triangle T} z_i\right] = \begin{cases} 1 & \text{if } S \triangle T = \emptyset \\ 0 & \text{if } S \triangle T \neq \emptyset \end{cases} \qquad (z_i \text{ is } +1 \text{ and } -1 \text{ with equal prob for any } i \in S \triangle T)$$

---

Above inner product $\langle \cdot, \cdot \rangle$ induces (Euclidean) **norm** $\|f\| = \sqrt{\langle f, f\rangle}$
Since every $f : X \to \mathbb{R}$ has unique expansion $f = \sum_{S\subseteq[n]} a_S c_S$ in Fourier basis,

$$\|f\|^2 = \langle f, f\rangle = \left\langle \sum_{S\subseteq[n]} a_S c_S, \sum_{T\subseteq[n]} a_T c_T \right\rangle = \sum_{S,T\subseteq[n]} a_S a_T \langle c_S, c_T\rangle = \sum_{S\subseteq[n]} a_S^2 \qquad \text{(Parseval theorem)}$$

Last equality uses orthonormality of Fourier basis
Coefficient $a_S = \langle f, c_S\rangle$ because $\langle f, c_S\rangle = \left\langle \sum_{T\subseteq[n]} a_T c_T, c_S \right\rangle = \sum_{T\subseteq[n]} a_T \langle c_T, c_S\rangle = a_S$

---

**Theorem 1.** *Let $\mathcal{D}$ be the uniform distribution over $X = \{+1,-1\}^n$. Any algorithm for learning $\mathcal{C} = \{\text{parity functions}\}$ to error $\varepsilon < 1/2$ from statistical queries of tolerance $\tau$ must query* $\mathrm{STAT}(c, \mathcal{D})$ *at least $(4|\mathcal{C}| - \gamma^{-2})\tau^2$ times, where $\gamma = \frac{1}{2} - \varepsilon$*

Since $|\mathcal{C}| = 2^n, \tau \geqslant 1/\operatorname{poly}(n)$ and $\gamma \geqslant 1/\operatorname{poly}(n)$, Theorem implies #queries $\geqslant \exp(\Omega(n))$

*Proof.* Let $c_S \in \mathcal{C}$ be the target concept, and $\varphi_1, \dots, \varphi_T$ all the query predicates to $\mathrm{STAT}(c_S, \mathcal{D})$
Expand each predicate, say $\varphi : X \times \{+1,-1\} \to \{0,1\}$, as $\varphi(x, y) = f(x) + g(x)y$
Intuitively, only 2nd term $g(x)y$ depends on label $y$ and reveals information about $c_S$
Each query corresponds to estimating

$$P_\varphi = \mathop{\mathbb{E}}_{z\in\{+1,-1\}^n}[\varphi(x, c_S(x))] = \mathop{\mathbb{E}}_{z\in\{+1,-1\}^n}[f(x) + g(x)c_S(x)] = \mathop{\mathbb{E}}_{z\in\{+1,-1\}^n}[f(z)] + \langle g, c_S\rangle$$

Suppose $\mathrm{STAT}(c_S, \mathcal{D})$ always answers every statistical query $(\varphi, \tau)$ with response $\hat{P}_\varphi = \mathbb{E}_z[f(z)]$
In other words, the response says that $|P_\varphi - \hat{P}_\varphi| = |\langle g, c_S\rangle| \leqslant \tau$

After $T$ queries, algorithm outputs hypothesis $h : X \to \{+1,-1\}$

Will show that some $c_S \in \mathcal{C}$ consistent with all answers has $\mathrm{err}_{\mathcal{D}}(h, c) > \varepsilon$
Which $c_S \in \mathcal{C}$ are ruled out when algorithm knows $|\langle g, c_S\rangle| \leqslant \tau$?
Let $A = \{S \subseteq [n] \mid |\langle g, c_S\rangle| > \tau\}$

**Claim 2.** $|A| \leqslant \|g\|^2/\tau^2$

*Proof.* Let $g_A = \sum_{S\in A}\langle g, c_S\rangle c_S$ (projection of $g$ to span of those $c_S$ with large inner product)
By Parseval, $\|g_A\|^2 = \sum_{S\in A}\langle g, c_S\rangle^2 \leqslant \sum_{S\subseteq[n]}\langle g, c_S\rangle^2 = \|g\|^2$
On the other hand, $\|g_A\|^2 = \sum_{S\in A}\langle g, c_S\rangle^2 \geqslant |A|\tau^2$ by definition of $A$ □

$\|g\|^2 = \mathbb{E}_{z\in\{+1,-1\}^n}[g(z)^2] \leqslant 1/4$ because $|g(x)| = |(\varphi(x,1) - \varphi(x,-1))/2| \leqslant 1/2$
By Claim, at most $1/4\tau^2$ many $c_S \in \mathcal{C}$ are ruled out by a single response $|\langle g, c_S\rangle| \leqslant \tau$
After $T$ queries, at most $T/4\tau^2$ many parity functions are ruled out

How many $c_S \in \mathcal{C}$ has $\mathrm{err}_{\mathcal{D}}(h, c_S) \leqslant \varepsilon$?

$$\mathrm{err}_{\mathcal{D}}(h, c_S) = \mathop{\mathbb{P}}_{x\in\mathcal{D}}[h(x) \neq c_S(x)] = \frac{1 - \mathbb{E}_{x\in\mathcal{D}}[h(x)c_S(x)]}{2} = \frac{1 - \langle h, c_S\rangle}{2}$$

Define advantage $\gamma = \frac{1}{2} - \varepsilon$, then $\mathrm{err}_{\mathcal{D}}(h, c_S) \leqslant \varepsilon \iff \langle h, c_S\rangle \geqslant 2\gamma$
Again need to bound number of $c_S \in \mathcal{C}$ with large inner product with some function $h$
$\|h\|^2 = 1$ because $|h(x)| = 1$ for all $x \in X$
By calculations in Claim, at most $1/4\gamma^2$ many $c_S \in \mathcal{C}$ have $\mathrm{err}_{\mathcal{D}}(h, c_S) \leqslant \varepsilon$

If $\frac{T}{4\tau^2} + \frac{1}{4\gamma^2} < |\mathcal{C}|$, some $c_S \in \mathcal{C}$ consistent with all responses has $\mathrm{err}_{\mathcal{D}}(h, c) > \varepsilon$
Algorithm needs $\frac{T}{4\tau^2} + \frac{1}{4\gamma^2} \geqslant |\mathcal{C}| \implies T \geqslant (4|\mathcal{C}| - \gamma^{-2})\tau^2$ □