CS 224 Fall 2023 Scribes: Weiyuan Gong
Nov 8th

Lecture 18: Statistical Query Lower Bounds II

1 Overview

- Statistical query (SQ) model
- SQ lower bounds for supervised problems
  - SQ Lower bounds for noisy parity
  - SQ dimension and the general recipe for proving correlational SQ (CSQ) lower bounds
  - SQ Lower bounds for multilayer perceptrons (MLPs)
  - SQ Lower bounds for real-value functions
- SQ lower bounds for unsupervised problems
  - SQ lower bounds for unsupervised problems
  - How to lower bound statistical dimension

2 Statistical query (SQ) model

We define the statistical query (SQ) model of computation. We consider an algorithm that only interacts with dataset $\{(z_i)\}$ through some $\psi : \mathbb{R}^m \to \mathbb{R}$. The samples are in the forms of $\mathbb{E}[\psi(z)] + \zeta$, where ζ is some noise with $|\zeta| \leq \tau$ for tolerance τ corresponding to $\sqrt{1/N}$ (N is the number of samples). This model captures essentially any known learning algorithm except for Gaussian elimination.

3 SQ lower bounds for supervised problems

3.1 SQ Lower bounds for noisy parity

We start with the task of learning parity (LPN) with noise, which is a noisy supervised learning task. Given dimension d and parameter $\eta < 1$, our goal is to learn a

randomly picked $S \subset [d]$ given a data set $(x_1, y_1), ..., (x_N, y_N)$ as follows:

$$x \sim \{\pm 1\}^d, \qquad y = \begin{cases} x_S, & \text{w.p. } 1-\eta \\ - x_S, & \text{otherwise.} \end{cases}$$

We consider some upper bounds for this task:

- $N = d$ samples suffice information-theoretically (brute force over S).
- When $\eta = 0$, this is just a linear system modulo 2, can solve in polynomial time with Gaussian elimination.
- When $\eta > 0$, there is a $2^{O(d/\log d)}$-time algorithm that can beat brute force [BKW03].

We also have the following two properties from the lower bound side:

- **LPN hypothesis:** this task is hard to learn. Actually, there is no polynomial time algorithm even to even to distinguish from random labels.
- **k-sparse parity with noise:** if $|S| = k$, any algorithm requires $d^{\Omega}(k)$ time.

In particular, we have the following lower bound

Theorem 1 ([BFJ+94]). *Any statistical query algorithm for learning parity with noise requires $O(2^{\Omega(d)}$ queries or tolerance $2^{-\Omega(d)}$.*

proof strategy. We consider any query to $\mathbb{E}_{x,y}[y \cdot \psi(x)]$ and:

- Argue that as a random variable in the unknown parity S, this quantity concentrates around its expectation $\mathbb{E}_S \mathbb{E}_{x,y}[y \cdot \psi(x)]$.
- For most S, a valid SQ oracle would simply answer the query with $\mathbb{E}_S \mathbb{E}_{x,y}[y \cdot \psi(x)]$.
- This oracle provides very little information about S.

□

3.2 SQ dimension

Now, we provide General recipe for proving CSQ lower bounds for supervised problems: statistical query dimension using the so called SQ dimensions. We recall the correlational statistical query (CSQ) models. In this setting, the algorithm only interacts with the dataset through queries of the form $\mathbb{E}[y \cdot \psi(x)] + \zeta$, where $\psi : \mathbb{R}^d \to \mathbb{R}$ and ζ is some noise with $|\zeta| \leq \tau$ for tolerance τ corresponding to $\sqrt{1/N}$. We provide the definition for SQ dimension as follows:

Definition 1 (SQ dimension). *A class of functions has SQ dimension $\geq D$ w.r.t. input distribution q if there exist functions $f_1, ..., f_D$ in the class s.t. for all $i \neq j$:*

$$\left|\mathbb{E}_{x\sim q}[f_i(x)f_j(x)]\right| \leq \frac{1}{D}.$$

As an example, we consider the PARITY $= \{f(x) = x_S : S \subset [d]\}$ has SQ dimension 2^d with respect to uniform distribution. This argument also lets us generalize to sparse parity.

Now, we give the following theorem connecting the SQ dimension and CSQ query lower bounds.

Theorem 2 ([BFJ+94, Szö09]). *If $\mathcal{F}$ has SQ dimension D with respect to q, then any CSQ algorithm for learning $\mathcal{F}$ from examples from q requires $\Omega(D\tau^2)$ queries or tolerance τ.*

As an example, we take the tolerance to be $\tau = \sqrt[3]{1/D}$,. This indicates that If D is super-polynomially large, the SQ lower bound qualitatively implies that you either need a super-polynomial number of queries or a super-polynomial number of samples (inverse tolerance).

proof on board. We consider $\mathcal{F}$ with SQ dimension $\geq D$, i.e., we can find $f_1, ..., f_D \in \mathcal{F}$ such that

$$\left|\mathbb{E}_{x}[f_i(x)f_j(x)]\right| \leq \frac{1}{D}.$$

The brief idea for the proof is to argue that the number of problem instances $i \in [D]$ that get "ruled out" at every state is very small because for most i, answering with a "trivial" oracle response is accurate.

In particular, we define

$$\langle f, g\rangle := \mathbb{E}_{x\sim q}[f(x)g(x)].$$

Fixing CSQ query $\mathbb{E}[y\psi(x)]$, we further define

$$A^+ := \{i \in [D] : \langle f_i, \psi\rangle \geq \tau\},$$
$$A^- := \{i \in [D] : \langle f_i, \psi\rangle \leq -\tau\}.$$

Our goal is to show that $\left|A^{\pm}\right|$ are small. We pick the potential function

$$Z = \langle \psi, \sum_{i\in A^+} f_i\rangle^2.$$

We first provide the upper bound of Z by Cauchy-Schwartz inequality:

$$\begin{aligned}
Z &\leq \|\psi\|^2 \left\| \sum_{i \in A^+ f_i} \right\|^2 \\
&\leq \sum_{i,j \in A^+} \langle f_i, f_j \rangle \\
&= \sum_{i \in A^+} \|f_i\|^2 + \sum_{i \neq j \in A^+} \langle f_i, f_j \rangle \\
&\leq \sum_{i \in A^+} \|f_i\|^2 + \frac{1}{D} \left|A^+\right| \left(\left|A^+\right| - 1\right) \\
&\leq \frac{\left|A^+\right|^2}{D} + \left|A^+\right|.
\end{aligned}$$

We also have the following lower bound on Z according to the fat that $\langle \psi, \sum_{i \in A^+} f_i \rangle \geq \tau \left|A^+\right|$. By definition of A^+, we have

$$Z \geq \tau^2 \left|A^+\right|^2.$$

Therefore, we have

$$\tau^2 \left|A^+\right|^2 \leq Z \leq \frac{\left|A^+\right|^2}{D} + \left|A^+\right|,$$

which indicates that

$$\left|A^+\right| \leq \frac{D}{D\tau^2 - 1} \leq O(1/\tau^2).$$

Similarly, we have $\left|A^-\right| \leq O(1/\tau^2)$. This shows that all but $O(1/\tau^2)$ are consistent with the answer 0. Therefore, $D\tau^{-2}$ queries are enough to narrow down to the true answer. □

3.3 SQ Lower bounds for MLPs

For MLPs, we have the following lower bound for any CSQ algorithms.

Theorem 3 ([DKKZ20]). *Any CSQ algorithm for learning one-hidden-layer size-k MLP's over Gaussians, even to constant error, requires $2^{d^{\Theta(1)}}$ queries or tolerance $d^{-\Omega(k)}$*

In this class, we prove a slightly weaker bound with a simpler proof due to [GGJ+20]. In particular, they exhibit a family of $d^{\Omega(\log k)}$ networks that are all exactly orthogonal to each other.

Proof. We select $S \subseteq [d]$ of size $m = \log k$. Given $w \in \{\pm 1\}^m$, we defined $w^{[S]} \in \mathbb{S}^{d-1}$ by

$$w_i^{[S]} = \begin{cases} w_i/\sqrt{m}, & \text{if } i \in S; \\ 0, & \text{otherwise.} \end{cases}$$

We define

$$f_S(x) = \sum_{w \in \{\pm 1\}^m} (\prod_{i=1}^{m} w_i) \text{ReLU}(\langle w^{[S]}, x \rangle).$$

We have the following two claims:

Claim 1. *f_S is nonzero.*

Claim 2. *$\langle f_S, f_T \rangle = 0$ for any sign-symmetric q if $S \neq T$.*

Proof. We define $\odot$ as $(x \odot z)_i = x_i z_i$. We have

$$\begin{aligned} f(x \odot z) &= z_S f_S(x) \\ &= \sum_w \prod_{i \in [m]} w_i \text{ReLU}(\langle w^{[S]}, x \odot z \rangle) \\ &= \sum_w \prod_{i \in [m]} w_i \text{ReLU}(\langle w^{[S]} \odot z, x \rangle) \\ &= \sum_{w'} \prod_{i \in [m]} w'_i \cdot z_S \text{ReLU}(\langle w'^{[S]}, x \rangle). \end{aligned}$$

Thus, we have

$$\begin{aligned} \langle f_S, f_T \rangle_q &= \mathbb{E}_x[f_S(x) f_T(x)] \\ &= \mathbb{E}_{x,z}[f_S(x \odot z) f_T(x \odot z)] \\ &= \mathbb{E}_{x,z}[f_S(x) f_T(x) z_S z_T] \\ &= \mathbb{E}_x[f_S(x) f_T(x)] \cdot \mathbb{E}_z[z_S z_T]. \end{aligned}$$

We can observe that the second term is 0, which makes $\langle f_S, f_T \rangle_q = 0$. □

Using the above two claims, we can deduce that the SQ dimension for this problem is at least

$$\text{SQ}_{d,m} \geq d^{\Omega(m)} = d^{\Omega(\log k)}.$$

□

3.4 SQ Lower bounds for real-value functions

The Full SQ lower bounds for supervised learning of real-valued functions are rare or hard to show. This is because the quirk of the SQ model. However, we still have the following observation:

Observation 1 ([VW19]). *Suppose $\mathcal{F}$ is a finite collection of functions such that for every $f, g \in \mathcal{F}$, $\Pr_{x\sim q}[f(x) = g(x)] = 0$. Then there exists a statistical query that will rule out a constant fraction of functions in $\mathcal{F}$, even with tolerance 0.1.*

proof strategy. The intuition is to find $\phi(x, f(x))$ such that $\phi(x, f(x)) = v_f$ for $v_f \in [0, 1]$. This is well-defined because for every $f, g \in \mathcal{F}$, $\Pr_{x\sim q}[f(x) = g(x)] = 0$. In this way, an answer to $\phi(x, f(x))$ rules out all $g \in \mathcal{F}$ such that $\left|v_g - v_f\right| > 0.1$. □

We remark that when functions are real-valued but take on Boolean values a non-vanishing fraction of the time, then the proof above does not apply. Other than this observation, [CGKM22] also provides a full SQ lower bounds for learning (real-valued) MLPs over Gaussian inputs

4 SQ lower bounds for unsupervised problems

4.1 SQ lower bounds for unsupervised problems

Instead of families of functions w.r.t. some input distribution q, we now consider families of distributions. Let D be a reference distribution (typically a simple distribution like Unif$(\{\pm 1\}^d)$ or $\mathcal{N}(\mathbf{0}, I_d)$. We define pairwise correlation between two distributions relative to D as

$$\langle D_1, D_2\rangle_D = \mathbb{E}_{x\sim D}\left[\left(\frac{D_1(x)}{D(x)} - 1\right)\left(\frac{D_2(x)}{D(x)} - 1\right)\right].$$

Given set T of distributions, we define average correlation w.r.t. D by

$$\rho_D(T) = \frac{1}{|T|^2}\sum_{D_1,D_2\in T}\langle D_1, D_2\rangle_D.$$

In particular, we say the set of distributions T^* has statistical dimension $\geq \Delta$ with respect to D with average correlation γ if for every $T \subseteq T^*$ of size $\geq |T^*|/\Delta$, we have $\rho_D(T) \leq \gamma$. We have the following theorem:

Theorem 4 ([FGR$^+$17]). *Suppose T^* is a finite collection of distributions with statistical dimension $\geq \Delta$ with respect to D with average correlation γ. Then any statistical query algorithm for learning distributions in T^* requires tolerance $\sqrt{\gamma}$ or at least $\Omega(\Delta)$ queries*

proof intuition. Suppose that for $> 1/\Delta$ fraction of distributions in T^*, some statistical query has expectation much farther than its expectation w.r.t. D. Then the average correlation among those distributions is too large. The remaining proof closely tracks the SQ dimension proof for supervised learning. □

4.2 How to lower bound statistical dimension

In this subsection, we focus on deriving a general recipe for SQ lower bounds. We start from an SQ lower bound for mixtures of Gaussians. Our goal is to design a set of Gaussian mixtures which mostly have tiny pairwise correlation with each other relative to $\mathcal{N}(0, I_d)$.

A core problem here is to get tiny pairwise correlation. The idea here is the so-called "moment matching". In particular, we consider some distribution A over $\mathbb{R}$ that matches the first m order moments with $\mathcal{N}(0, I_d)$. We then define P_v to be the moment of unit vector v w.r.t. A. These P_v's can be regarded as some "parallel pancakes" that satisfy the two key properties:

- Moments of P_v are equal to moments of $\mathcal{N}(0, I_d)$.
- For typical v', the projection of P_v along v' looks like $\mathcal{N}(0, I_d)$.

Therefore, the method of moments and the dimensionality reduction (actually one can prove that for all CSQ algorithms) fail. Formally, we have the following theorem:

Theorem 5 ([DKS17]). *Let A be a distribution over $\mathbb{R}$ whose first m moments match those of $\mathcal{N}(0, 1)$, i.e.*

$$\mathbb{E}_{x\sim A}[x^i] = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[x^i]$$

for all $1 \le i \le m$. Then for any unit vector u, v, we have

$$\langle P_u, P_v\rangle_{\mathcal{N}(0,I_d)} \le \left|\langle u, v\rangle\right|^{m+1} \langle A, A\rangle_{\mathcal{N}(0,1)}.$$

From the mixture of Gaussian, we propose the general recipe for SQ lower bounds:

- We construct one-dimensional moment-matching example using a distribution from the distribution family in question (e.g. Gaussian mixtures). This step is usually highly problem-specific and where all the hard work goes.
- We hide it along some direction (P_v should still be a member of the distribution family).

- We argue that it is hard for any SQ algorithm to distinguish whether samples come from some P_v or from $\mathcal{N}(0, I_d)$.

In conclusion, we have the following theorem for learning mixture of Gaussians:

Theorem 6 ([DKS17]). *Any SQ algorithm that learns general mixtures of Gaussians, i.e. of the form $\sum_{i=1}^{k} \lambda_i \mathcal{N}(\mu_i, \Sigma_i)$, requires $d^{\Omega(k)}$ queries or $d^{-\Omega(k)}$ tolerance.*

References

[BFJ+94] Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 253–262, 1994.

[BKW03] Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)*, 50(4):506–519, 2003.

[CGKM22] Sitan Chen, Aravind Gollakota, Adam Klivans, and Raghu Meka. Hardness of noise-free learning for two-hidden-layer neural networks. *Advances in Neural Information Processing Systems*, 35:10709–10724, 2022.

[DKKZ20] Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539. PMLR, 2020.

[DKS17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

[FGR+17] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2):1–37, 2017.

[GGJ+20] Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pages 3587–3596. PMLR, 2020.

[Szö09] Balázs Szörényi. Characterizing statistical query learning: simplified notions and proofs. In *International Conference on Algorithmic Learning Theory*, pages 186–200. Springer, 2009.

[VW19] Santosh Vempala and John Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds. In *Conference on Learning Theory*, pages 3115–3117. PMLR, 2019.