Lecture 28: Statistical Query Algorithms

Daniel Kane
Scribe: Ken Hoover

June 5, 2017

Statistical Query Model

So far, we've given a number of algorithms for various problems in the robust statistics setting. What we would now like are lower bounds on the sample complexity of these problems. In particular, we would like lower bounds for computationally efficient algorithms. This proves difficult to do with purely information theoretic tools, as we've been using thus far; while we can prove lower bounds, the algorithms matching those bounds take exponential amounts of time to run.[1] But with information theoretic techniques not being useful, it becomes hard to talk about algorithms with arbitrary access to samples.

What we then want to do is restrict the access an algorithm has to samples, in a reasonable way, and then prove lower bounds given that restriction. Note that many of the algorithms so far seem to just be computing averages of various functions over the distribution. A reasonable restriction, then, might be only allowing access to the distribution via what the expected value of some function over the distribution is.

To put this formally, say we have a distribution X over some domain $\mathcal{D}$. Let $f : \mathcal{D} \to [0,1]$ be a bounded function.[2] We define an oracle STAT(ϵ) which, given such an f, returns a value in $[\mathbb{E}_{x\sim X}[f(x)] - \epsilon, \mathbb{E}_{x\sim X}[f(x)] + \epsilon]$. Note that N calls to STAT(ϵ) can be simulated with $\log N/\epsilon^2$ samples and N time. This simulation is also robust; up to an $\epsilon/2$ fraction of adversarial noise can be present while still obtaining correctness. This result is even better in the weaker additive error model, where an arbitrarily large constant fractions of samples can be from a known noise distribution. Even better, the filter-based algorithms used so far can be transformed to fit into this model.

Statistical Query Lower Bounds

Learning Parities

Our first use of the SQ model for lower bounds will be the following. Suppose we have a distribution X on $\mathbb{F}_2^n$ that is the uniform distribution over a random subspace of codimension 1. Our goal is to learn X. We note that, outside of the SQ model, we would take n samples, perform Gaussian Elimination on the resulting matrix, and have an algorithm for finding the subspace.

[1] Or are NP-hard.

[2] Any bounded function over the reals is isomorphic to one over $[0,1]$.

For this task, SQ on its own is insufficient for proving bounds. Thus, we bring in Fourier Analysis. As a brief reminder, the Fourier transform $\widehat{f} : 2^{[n]} \to \mathbb{R}$ of f is the unique function such that

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S)\chi_S(x) \ ,$$

where $\chi_S(x) = \prod_{i \in S}(-1)^{x_i}$. The two properties of the Fourier transform we shall be using are the following:

- **Plancherel's Theorem:** For any $f : \mathbb{F}_2^n \to [0, 1]$,

  $$||\widehat{f}(S)||_2^2 = \mathbb{E}_{x \sim \mathcal{U}}[|f(x)|^2] \leq 1 \ ,$$

  where $\mathcal{U}$ is the uniform distribution over $\mathbb{F}_2^n$.

- If a distribution X is uniform over $\{x : \chi_S(x) = 1\}$ for some $S \subseteq [n]$, then

  $$\mathbb{E}_{x \sim X}[f(x)] = \widehat{f}(\emptyset) + \widehat{f}(S) \ .$$

An important consequence of Plancherel's Theorem is that for almost all subsets $S \subseteq [n]$, $|\widehat{f}(S)|$ is small. In particular, for any given ϵ, we have that at most $1/\epsilon^2$ subsets S exist for which $\widehat{f}(S) > \epsilon$. Thus, for any given f an oracle can, with high probability, just return $\widehat{f}(\emptyset)$ so long as $\epsilon >> 2^{-n/2}$. This results in very little being learned about X; this can be formalized as saying that, under this distribution, STAT(()ϵ) has exponentially small (in n) entropy. So, we end up with an $\Omega(2^n\epsilon^2)$ query lower bound to STAT(()ϵ); moreover, if we have exponentially accurate queries, then simulating the oracle also takes exponential time.

Generalization to Orthogonal Distributions

In the previous section, we needed to pick one distribution out of a large set of possible options. These distributions were all, in some sense, nearly orthogonal to each other. We now want to see if these lower bounds continue to hold when general distributions are orthogonal; to do this, we need to define an inner product on distributions. Let $\mathcal{P}$, $\mathcal{Q}$, and $\mathcal{D}$ be distributions over some domain X. We define the inner product of $\mathcal{P}$ and $\mathcal{Q}$ as follows:

$$\langle \mathcal{P}, \mathcal{Q} \rangle_{\mathcal{D}} = \sum_{x \in X} \frac{\mathcal{P}(x)\mathcal{Q}(x)}{\mathcal{D}(x)} \qquad \left(= \int_X \frac{\mathrm{d}\mathcal{P} \cdot \mathrm{d}\mathcal{Q}}{\mathrm{d}\mathcal{D}} \text{ for continuous domains} \right) \ .$$

An issue with this definition is that $\langle \mathcal{P}, \mathcal{Q} \rangle_{\mathcal{D}} \geq 1$, so we shall instead use

$$\chi_{\mathcal{D}}(\mathcal{P}, \mathcal{Q}) = \langle \mathcal{P} - \mathcal{D}, \mathcal{Q} - \mathcal{D} \rangle_{\mathcal{D}} = \langle \mathcal{P}, \mathcal{Q} \rangle_{\mathcal{D}} - 1 \ .$$

Suppose we have a distribution $\mathcal{P} \in \mathcal{C}$ we are trying to learn, where $\mathcal{C}$ is a finite set of distributions. Also suppose that there is some distribution $\mathcal{D}$, and a set of distributions $\{\mathcal{P}_1, \ldots, \mathcal{P}_N\} \subseteq \mathcal{C}$ such that

$$|\chi_{\mathcal{D}}(\mathcal{P}_i, \mathcal{P}_j)| < \delta \ \ \forall i \neq j,$$

and

$$\chi_{\mathcal{D}}(\mathcal{P}_i, \mathcal{P}_i) < C \ \ \forall i \text{ and some } C.$$

We want to show that for any query function $f(x)$,

$$\mathbb{E}_{x\sim\mathcal{P}_i}[f(x)] \approx \mathbb{E}_{x\sim\mathcal{D}}[f(x)]$$

for most $\mathcal{P}_i$. Note that the former is equal to $\langle \mathcal{P}_i, f\circ\mathcal{D}\rangle_{\mathcal{D}}$, while the latter is equal to $\langle \mathcal{D}, f\circ\mathcal{D}\rangle_{\mathcal{D}}$; so all we need to do is show that $\langle \mathcal{P}_i - \mathcal{D}, f\circ\mathcal{D}\rangle_{\mathcal{D}}$ is small.

To do this, we first observe that

$$\langle f\circ\mathcal{D}, f\circ\mathcal{D}\rangle_{\mathcal{D}} = \int \frac{f^2 \mathrm{d}\mathcal{D}^2}{f^2 \mathrm{d}\mathcal{D}} = \int f^2 \mathrm{d}\mathcal{D} \leq 1 \ .$$

Suppose that $|\langle f\circ\mathcal{D}, \mathcal{P}_{i_j} - \mathcal{D}\rangle_{\mathcal{D}}| > \epsilon$ for $j = 1, \ldots, m$. Consider $g = f\circ\mathcal{D} - x(\sum_{j=1}^m \pm(\mathcal{P}_{i_j} - \mathcal{D}))$. We then have that

$$\begin{aligned} 0 \leq \langle g, g\rangle_{\mathcal{D}} &= \langle f\circ\mathcal{D}, f\circ\mathcal{D}\rangle_{\mathcal{D}} - 2x\sum_{j=1}^m \langle f\circ\mathcal{D}, \pm\mathcal{P}_{i_j} - \mathcal{D}\rangle_{\mathcal{D}} \\ &\quad + x^2\left[\sum_{j=1}^m \langle \mathcal{P}_{i_j} - \mathcal{D}, \mathcal{P}_{i_j} - \mathcal{D}\rangle_{\mathcal{D}} + \sum_{1\leq j<k\leq m} \pm\langle \mathcal{P}_{i_j} - \mathcal{D}, \mathcal{P}_{i_j} - \mathcal{D}\rangle_{\mathcal{D}}\right] \\ &\leq 1 - 2xm\epsilon + x^2(mC + m^2\delta) \ . \end{aligned}$$

Note that this is positive for every value of x, so we have that

$$m\epsilon < \sqrt{mC + m^2\delta} \ ,$$

or

$$\epsilon < \sqrt{C/m + \delta} \ .$$

In particular, if $m > C/\delta$, then $\epsilon < \sqrt{2\delta}$. Thus, the difference between $\mathbb{E}_{x\sim\mathcal{P}_i}[f(x)]$ and $\mathbb{E}_{x\sim\mathcal{D}}[f(x)]$ is greater than $\sqrt{2\delta}$ for at most C/δ many $\mathcal{P}_i$s. Thus, learning $\mathcal{P}$ requires $N\delta/C$ queries to a STAT($\sqrt{2\delta}$) oracle. This either requires an exponential amount of time, or a large number of samples to get the required accuracy.

As it turns out, this is the only obstacle to learning $\mathcal{P}$ efficiently. Suppose $\mathcal{C} = \{\mathcal{P}_i, \ldots, \mathcal{P}_N\}$. The SQ-dimension of $\mathcal{C}$ is the maximum integer d such that there is a distribution $\mathcal{D}$ and a subset $\{\mathcal{Q}_1, \ldots, \mathcal{Q}_d\} \subseteq \mathcal{C}$ such that

$$\chi_{\mathcal{D}}(\mathcal{Q}_i, \mathcal{Q}_i) \leq d$$

and

$$|\chi_{\mathcal{D}}(\mathcal{Q}_i, \mathcal{Q}_j)| \leq 1/d$$

for every $i \neq j$. We shall show in the next lecture that if the SQ-dimension of a class is d, then there is an SQ algorithm which uses $\mathrm{poly}(d)\log N$ queries to a STAT($\mathrm{poly}(1/d)$) oracle which learns any distribution in $\mathcal{C}$ to within $1/\mathrm{poly}(d)$.