Lecture 17: Structured Distributions: Lower Bound on Learning and Property Testing

Daniel Kane
Scribe: Nadim Ghaddar

Lecture on May 10, 2017

During the last lecture, we got to know that learning a distribution p that is ϵ-approximately t-piecewise, degree d polynomial up to total variation distance ϵ requires $\tilde{O}(\frac{t(d+1)}{\epsilon^2})$ samples, where $\tilde{O}$ hides logarithmic factors. In this lecture, we will show that this is actually a lower bound as well. And then we move on to the topic of property testing for structured distributions.

1 Lower Bound on Learning Structured Distributions

Theorem 1. *Any algorithm which learns a t-piecewise degree-d distribution up to error ϵ requires $\Omega(\frac{t(d+1)}{\epsilon^2})$ samples.*

Remember that learning an unstructured distribution over $[n]$ required $\theta(\frac{n}{\epsilon^2})$ samples. The idea of the proof here is to show that for a t-piecewise degree-d distribution, there are $t(d+1)$ "almost-independent" values for the probability density function, and thus this is equivalent to learning an unstructured distribution with $t(d+1)$ bins.

For a t-piecewise degree-d distribution, we know we have t bins, and thus doing some independent functions on each bin will result in the factor of t in the lower bound. But the idea is to show that on each bin, a minimum of d independent values are needed. For that, we consider Chebyshev polynomials on $[-1, 1]$:

$$P(\cos(\theta)) = \frac{(\sum_{k=0}^{d} \cos(k(\theta - \phi)))^2 + (\sum_{k=0}^{d} \cos(k(\theta + \phi)))^2}{d^2}$$

which are symmetric around ϕ, highly concentrated in the interval $\theta \in (\phi \pm 1/d)$ and falls off quadratically otherwise.

Let P_m be the normalization of P centered at $\cos(\frac{2\pi m}{d})$. Note that there are d different P_m's of degree $= O(d)$ and that overlap very little.

To learn the distribution p, we follow an adversarial method as usual. We define p on each of the t intervals as follows:

$$p = \frac{\frac{1}{d}\sum_{m=0}^{d-1}(1 \pm 10\epsilon)P_m(x)}{t}$$

The point is that if we want to learn p up to error ϵ, we need to guess $\frac{2}{3}$ of the the plus/minus signs correctly. The basic point of the proof is that for the t-piecewise

degree-d distribution, there are t intervals, and on each interval, it is piecewise polynomial. The distribution p is the sum of the P_m functions multiplied by either $(1+10\epsilon)$ or $(1-10\epsilon)$, but since the P_m functions fall off quadratically away from the center, then the main contribution in each sub-interval comes from the primary function. Therefore, errors mostly rise from getting the sign change wrong for the primary function in each sub-interval. In other words, learning p up to error ϵ requires learning $\frac{2}{3}$ of the number of signs in the $t(d+1)$ sub-intervals correctly. The rest of the argument follows from standard adversarial method techniques that we used before. If X^N is the string of signs and N the number of samples, then the shared information between X^N and the samples must be larger than $\frac{2}{3}$ of the number of sub-intervals (i.e. number of P_m functions), i.e. we should have

$$\begin{aligned} t(d+1) &\ll I(X^N;\ \text{samples}) \\ &\leq NI(X;\ \text{one sample}) \\ &\leq NO(\epsilon^2) \end{aligned}$$

where last inequality follows from the fact that the probability that a single sample ends up at a given value only varies by $(1 \pm \epsilon)$. And therefore, we should have $N \gg \frac{t(d+1)}{\epsilon^2}$.

2 Property Testing

Here we are given samples from a structured distribution p, and we need to distinguish between two possible cases: either $p = q$ or $d_{TV}(p,q) > \epsilon$. Similar to the learning problem, we can get bounds on the sample complexity of the property testing problem by using the A_k distance. Recall that

$$|p-q|_{A_k} = \sup_{\text{partition } \mathbb{R} \text{ into } I_1,\ldots,I_k} \frac{1}{2} \sum |p(I_i) - q(I_i)|$$

Idea: We need to find k such that for any p,$q \in \mathcal{C}$, we have

$$d_{TV}(p,q) = |p-q|_{A_k} \pm \frac{\epsilon}{2}$$

If this holds then it is enough to test $p = q$ vs $|p-q|_{A_k} > \frac{\epsilon}{2}$.

A useful analogy is the following: As we know, learning a distribution on $[n]$ requires $\frac{n}{\epsilon^2}$ samples. However, learning a distribution on $\mathbb{R}$ with respect to total variational distance requires infinite number of samples, whereas learning it with respect to A_k metric requires $\frac{k}{\epsilon^2}$ samples. This is because you can assume the distribution was a histogram over the k pieces and hence learning it is equivalent to learning an unstructured distribution with k. It turns out actually that $\frac{k}{\epsilon^2}$ samples is actually sufficient. On the other hand, identity testing for example on $[n]$ requires $\frac{\sqrt{n}}{\epsilon^2}$ samples, so we might expect by analogy that identity testing with respect to A_k metric requires something like $\frac{\sqrt{k}}{\epsilon^2}$ samples. By a similar argument

as for the learning case, it is similar that at least $\frac{\sqrt{k}}{\epsilon^2}$ samples are required. In what follows, we try to show that actually this number of samples is actually sufficient for identity testing.

Algorithm

Assuming $q \sim \mathcal{U}(0,1)$, the problem is to distinguish $p = q$ from $|p-q|_{A_k} > \epsilon$. We assume that q to be the uniform distribution because we can always find a unique change of variables that preserves the ordering on the x-axis and turns the distribution into a uniform distribution over $[0,1]$ (e.g. we can use the CDF of q as the new variable). Since the ordering on the x-axis is preserved then the A_k distance is not changed and the reduction is valid.

Another assumption is made, which is that the distribution p is a k-flat distribution. The A_k distance is exactly the variational distance in this case, and we know that there exists a partition into k intervals that gives us all we need to know about p, i.e. its A_k distance with respect to q.

If S is a partition into intervals, consider p^S and q^S to be the distributions induced by p and q respectively in the intervals that we land in. So we have that

$$|p-q|_{A_k} = \sup_{|S|=k} d_{TV}(p^S, q^S)$$

So the problem reduces to distinguishing, for any given S, between $p^S = q^S$ and $|p^S - q^S| > \epsilon$, which, as we know, requires $\sqrt{k}/\epsilon^2$ samples. However, the only remaining problem is to identify the partition to consider. If we consider S to be the partition into k equal intervals, it could be that p takes two values above and under the uniform distribution q for $k/2$ of intervals and agrees with q on the remaining intervals. The A_k distance in this case is pretty substantial. To solve this, we can consider a finer partition, as shown in the following algorithm:

Algorithm: If S_m is the partition into $k2^m$ equal parts, where $m = 0, \ldots, \log(1/\epsilon)$, we run testers on

$$p^{S_m} = q^{S_m} \text{ vs } |p^{S_m} - q^{S_m}|_2^2 > \frac{\epsilon^2 2^{2m/3}}{k}$$

with error probability $\leq \frac{1}{10*2^m}$. If the testers return $p^{S_m} = q^{S_m}$ for all m, then we return $p = q$. The analysis of the algorithm is continued in the next lecture, and we will try to generalize for the case when p is not a k-flat distribution.