CSE 291 Scribe Notes Lecture 19

Xiaomeng Wang

May 22, 2017

1 Previous Lecture

Last time we figured out how to do identity testing with respect to the A_k distance. If we find some partitions of intervals into m bins, where when we restrict those intervals, q^I will have L2 norm of $O(1/\sqrt{m})$, close to uniform on that partition. Then we can distinguish $p = q$ from $|p^I - q^I|_{A_k} > \epsilon$ with $O(\sqrt{k}/\epsilon^2)$ samples, which is optimal.

If we know what q is explicitly, then we can just pick I to be some nice partition into k/ϵ equal size bins for q, and then if $|p - q|_{A_k} > \epsilon$, then even the restriction to this partition will still be bigger than $\epsilon/2$ and the same algorithm works.

If we don't know q ahead of time, for unstructured identity testing, it was $\sqrt{n}/\epsilon^2)$, we expect the right answer to become $\sqrt{k}/\epsilon^2)$, which made it to be correct. The lower bound is $\sqrt{k}/\epsilon^2)$, the partition on the A_k distance is giving us the best value. Even without the partition, we are able to do this by taking a bunch oblivious partitions and doing different tests on each of them and that actually giving it to work. For closeness testing, we also had this lower bound $n^{2/3}/\epsilon^{4/3}$ and we expect this to become $k^{2/3}/\epsilon^{4/3}$, which is not actually correct in general.

2 Algorithm Approaches

1. Try to get some partition I by taking samples. Take poi(m) samples from q and use those samples as interval boundaries, which will end up approximately m intervals. We can show that $|q^I|_2 = O(1/sqrtm)$. However, we also need that once we restrict those intervals, $|p^I - q^I|_{A_k} > \epsilon$, especially when m « k.

2. Take poi(m) samples from (p+q)/2, i.e. sample from a random p and q. Sort the samples and keep track of which distribution they came form.
   Idea:
   If there is a big A_k distance, there are some reasonable size intervals like the following graph, and on these intervals, If you got two samples from the same interval, it's likely that they both come from p or same distribution. The fact that we got this cluster that p is more likely over q means that if we find two samples that are close to each other, it's more likely that they come from the same distribution than come from different distribution.

   Define statistic Z = # pairs of consecutive samples from same distribution - # pairs of consecutive samples from different distribution. For example, if samples come from *ppqpqqp*, Z = -2.

3 Behavior of Z

3.1 p = q

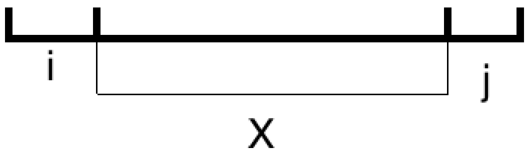
Sequence of p' s and q's, we get a uniform random sequence of length poi(m). So $Z \sim B'(max(0, poi(m) - 1))$ wp equals to $O(\sqrt(m))$ where B' is balanced binomial and $B'(n) = \sum_n \pm 1$.

Simplification: assume $p\&q$ are both continuous distributions, such that pdf(p+q) = 1 on [0,2]. The first thing to do is to remove atoms,then make a change of variables by $\sum$ of CPFs of p and q.

In order to further understand statistic Z, we can show that Var(Z) = O(m). To make argument vigorous, split [0,2] to m equal bins, Let $Z_i = \sum Z$ where I only consider pairs w/ first element of the pair is in i^{th} bin, so $Z = \sum Z_i$ and $Var(Z) = \sum cov(Z_i, Z_j)$, we need to figure out how that relates.

Lemmas for this to work:

- Var(Z_i) = O(1) because E[(♯ of samples in $i^{th} bin)^2$] = O(1), as ♯ of samples in i^{th} bin is distributed as poi(1).
- $Cov(Z_i, Z_j) = O(1)exp(-\Omega|i - j|)$.



If there is a bunch of bins between i and j, note that each of the bins have a constant probability of sample in them, $Cov(Z_i, Z_j|$samples from X) is going to be zero. Because anything lands in j will see the only sample come after j to determine how big z_j is, and z_i will see interacts between pairs within I and wherever the last thing in i is might interact with whatever comes next , but if wherever comes next is not in i, then there are no correlation of comes before j.So $\sum i = O(m)$

- $E[Z] = \int_0^2 f(t)(m/2)(dp - dq)$, where f(t) = prob(sample previous to t was from p) - prob(sample previous to t was from q).
When f(t) is defined as above, we can tell:
  - f(0) = 0
  - f(t + dt) = f(t)(1-(m/2)(pdt + qdt + O(dt^2)) + (m/2)(pdt -qdt + O(dt^2))) where dt is very small amount
  - f'(t) = f(t)(-(m/2)(p(t) + q(t)) + (m/2) (p(t) - q(t))) where p(t), q(t) are pdf for p,q
  - Rewrite the equation: $(m/2)(dp - dq) = f'(t)dt + (m/2)f(t)(dp + dq)$
  - $E[Z] = \int_0^2 f(t)f'(t)dt + (m/2)\int f^2(t)(dp + dq)$ where $f'(t) = f^2/2|_0^2 = O(1)$

Summarize:

- E[Z] = m/2 $\int f^2(t)$(dp + dq) + O(1)
- f'(t) = -(m/2) f(t)(p+q) + (m/2)(p-q)

We are in this case where p and q have reasonably large A_k distance, we want to know f is reasonably large, suppose interval I such that p(I) - q(I) = δ which is large.

WTS: f(t) is large somewhere on interval I

Pf: Introduce |I| = p(I) + q(I)

Look at f(I_{max}) - f(I_{min}) = $\int_I$ f'(t)dt = $\int_I$(m/2)f(t)(dp+dq) + $\int_I$(m/2)f(t)(dp-dq)
where $\int_I$(m/2)f(t)(dp+dq) is bounded by m/2 $|f|_\infty$ |I| and $\int_I$(m/2)f(t)(dp-dq) = mδ/2.
mδ/2 = O(((m/2)|I| + 1))$|f|_\infty$)
In other words, $|f|_\infty$ » (mδ/2)/((m/2)|I| + 1)
If |I| « 1/m, => $|f|_\infty$ » mδ/2

So whenever you have a interval that is not too long, on which you see a reasonably discrepancy between p and q, then you are going to see there is at least some point on that interval, where f is reasonably large.

But we don't only want f to be reasonably large, we want this interval to be reasonably large. If we make same assumption, assume |I| « 1/m, then (m/2)$\int_I f^2(t)$(dp+dq) » $m^3\delta^3$

Assume k>m, if $|p-q|_{A_k} > \epsilon$, partition into O(k) intervals I_i of length $<$ 1/k which is now $<$ 1/m St if we define delta the same way, $\epsilon\delta_i > \epsilon => \epsilon\delta^3 >=$ k$(\epsilon/k)^3 = \epsilon^3/k^2 =>$ E[Z] $>= m^3\epsilon^3/k^2$

In order to distinguish p = q and p far from q, we need

1. $m^3\epsilon^3/k^2 >> m^{1/2}$
2. $m >> k^{4/5}/\epsilon^{6/5}$

If |I| $>$ 1/m, run two testers, Z tester and partition into I and test whether $|p^I - q^I|_{A_k} > \epsilon$, we have optimal partition I which measures A_k distance.

- If we have intervals length $<$ 1/m, this contributes to Z, Z tester will be sufficient.
- If we have intervals length $>$ 1/m, replace I with I' which are small intervals, then either $\Delta(I')$approximately equals $\Delta(I) =>$ I contributes to $|p^I - q^I|_{A_k}$ tester, else discrepancy come largely from small intervals at ends, and this contributes to Z.