Lecture 18: Analysis of the Algorithm of Property Testing for Structured Distributions

Qiushi Huang

May 20, 2017

1 Previous Lecture

Last time, we studied the property testing for structured distribution with respect to A_k distance: $p, q \in C$, q is an explicitly given distribution, we want to distinguish between $p = q$ vs. $|p - q|_{A_k} > \epsilon$ given samples from distribution p.

Since we know q, we can change variables that turns q into uniform distribution for convenience. We assume $q = U(0, 1)$

The algorithm is as following:

Let I_m be the partition of $[0, 1]$ into $k2^m$ equal intervals, for $m = 0, 1, \ldots, \log(1/\epsilon)+O(1)$. Run testers to distinguish $p^{I_m} = q^{I_m}$ vs. $|p^{I_m} - q^{I_m}|_2^2 > \frac{\epsilon^2}{100k2^{m/3}}$ with error probability $< \frac{1}{2^m 100}$ If all testers return $p^{I_m} = q^{I_m}$, we will return $p = q$, otherwise we will return $|p - q|_{A_k} > \epsilon$

2 Sample complexity

We first analyze the sample complexity of this algorithm.

Notice that the L_2 tester over the distribution on $[n]$ need $O(\frac{||q||_2}{\epsilon^2}\log(1/\delta))$ to distinguish $p = q$ vs. $||p-q||_2 > \epsilon$ with probability at least $1-\delta$. Therefore, for each m in our algorithm, $||q^{I_m}||_2 = \frac{1}{\sqrt{k2^m}}$, then the sample complexity is

$$O(||q^{I_m}||_2 \frac{k2^{m/3}}{\epsilon^2}\log(2^m 100)) \leq O(\frac{\sqrt{k}}{\epsilon^2} m 2^{-m/6}) \leq O(\frac{\sqrt{k}}{\epsilon^2})$$

For $m = 0, 1, \ldots, \log(1/\epsilon)+O(1)$, we can use the same samples, so the sample complexity of the algorithm is $O(\frac{\sqrt{k}}{\epsilon^2})$

3 Special case: p is k-flat

We then show that this algorithm is correct.

1. If $p = q$, then for each $m = 0, 1, \ldots, \log(1/\epsilon) + O(1)$, we have $p^{I_m} = q^{I_m}$. Then with high probability(error probability $\leq \sum_m \frac{1}{2^m 100} < 1/50$), all the testers will return $p^{I_m} = q^{I_m}$, the algorithm will return $p = q$.

2. If $|p-q|_{A_k} > \epsilon$, we will show that there exists m such that $|p^{I_m} - q^{I_m}|_2^2 > O(\frac{\epsilon^2}{k2^{m/3}})$. Then this tester will return $|p^{I_m} - q^{I_m}|_2^2 > O(\frac{\epsilon^2}{k2^{m/3}})$ with high probability, which follows that the algorithm will return $|p-q|_{A_k} > \epsilon$.

We first consider the special case that p is k-flat.

Since p is k-flat, there exists a partition of $[0,1]$ into k intervals, let the ith interval has length l_i, and $p - q = \Delta_i dx$ in the ith interval since $p-q$ is constant in each interval.

Then we have $\sum l_i \Delta_i = 0$, $\sum l_i |\Delta_i| > \epsilon$

Now we will do some reduction on those intervals.

1. Chopping up intervals longer than $1/k$

   Since the total length of the intervals is 1, then we will at most create k new intervals with length $1/k$.

   Then we have $O(k) \leq 2k$ intervals with $l_i \leq 1/k$

2. Throw out intervals with $l_i < \frac{\epsilon}{8k}$

   Let $\Delta_i' = max(0, \Delta_i)$, then $\sum l_i |\Delta_i| = 2 \sum l_i \Delta_i' > \epsilon$, $\sum l_i \Delta_i' > \epsilon/2$

   Note that $\Delta_i' \leq 1$, $l_i \Delta_i' \leq l_i$, if we throw out intervals with $l_i < \frac{\epsilon}{8k}$, then sum of those intervals $\sum_{l_i < \frac{\epsilon}{8k}} l_i \Delta_i' \leq 2k \times \frac{\epsilon}{8k} = \frac{1}{4\epsilon}$, it follows

$$\epsilon/2 < \sum l_i \Delta_i' = \sum_{l_i \geq \frac{\epsilon}{8k}} l_i \Delta_i' + \sum_{l_i < \frac{\epsilon}{8k}} l_i \Delta_i' \leq \sum_{l_i \geq \frac{\epsilon}{8k}} l_i \Delta_i' + \epsilon/4$$

   Therefore, $\sum_{l_i \geq \frac{\epsilon}{8k}} l_i |\Delta_i| \geq \sum_{l_i \geq \frac{\epsilon}{8k}} l_i \Delta_i' > \epsilon/4$.

   Thus, after throwing out intervals with $l_i < \frac{\epsilon}{8k}$, we still have

$$\sum_{l_i \geq \frac{\epsilon}{8k}} l_i |\Delta_i| > \epsilon/4$$

After the reduction, we can assume that $\frac{\epsilon}{8k} < l_i < \frac{1}{k}$ for all l_i.

Now we analyze $|p^{I_m} - q^{I_m}|_2^2$. For interval with length l_i, pick m such that $\frac{l_i}{4} \leq \frac{1}{k2^m} < \frac{l_i}{2}$. Consider the endpoints of I_m, since $\frac{1}{k2^m} < \frac{l_i}{2}$, there are at least two endpoints of I_m in l_i, which means there exists an interval of I_m which is completely contained on interval l_i. For this interval, it contributes $(\Delta_i \frac{1}{k2^m})^2$ to $|p^{I_m} - q^{I_m}|_2^2$. Note that $\frac{l_i}{4} \leq \frac{1}{k2^m}$, we have $|p^{I_m} - q^{I_m}|_2^2 > (\Delta_i \frac{l_i}{4})^2 = \frac{(\Delta_i l_i)^2}{16}$.

Therefore, for every interval l_i, there exists m such that $|p^{I_m} - q^{I_m}|_2^2 > \frac{(\Delta_i l_i)^2}{16}$. It follows that

$$\sum_m |p^{I_m} - q^{I_m}|_2^2 > \sum_{i=0}^{O(k)} \frac{(\Delta_i l_i)^2}{16}$$

On the other hand, we have

$$\sum_{i=0}^{O(k)} l_i |\Delta_i| > \epsilon/4$$

By Cauchy-Schwarz, we have

$$\sum_{i=0}^{O(k)} \frac{(\Delta_i l_i)^2}{16} \geq \frac{(\sum_{i=0}^{O(k)} l_i|\Delta_i|)^2}{16O(k)} = O(\epsilon^2/k)$$

Therefore,

$$\sum_m |p^{I_m} - q^{I_m}|_2^2 > O(\epsilon^2/k)$$

which means there exists m such that $|p^{I_m} - q^{I_m}|_2^2 > O(\frac{\epsilon^2}{k2^{m/3}})$.(Otherwise, if every m, $|p^{I_m} - q^{I_m}|_2^2 < O(\frac{\epsilon^2}{k2^{m/3}})$, then $\sum_m |p^{I_m} - q^{I_m}|_2^2 < O(\epsilon^2/k)\sum_m 2^{-m/3} < O(\epsilon^2/k)$)

Thus, if $|p-q|_{A_k} > \epsilon$, we have showed that there exists m such that $|p^{I_m} - q^{I_m}|_2^2 > O(\frac{\epsilon^2}{k2^{m/3}})$.

The analysis above does not need $q = U(0,1)$.We only need $|q^{I_m}|_2 = O(\frac{1}{\sqrt{k2^m}})$ for $m = 0, 1, \ldots, \log(1/\epsilon) + O(1)$.

4 General case

We now consider the general case that p is not k-flat. We will show that the same algorithm works.

We know that the the sample complexity of the algorithm is $O(\frac{\sqrt{k}}{\epsilon^2})$. If $p = q$, the analysis is the same as before.

Now we consider the case $|p-q|_{A_k} > \epsilon$.

We can partition $[0,1]$ into intervals $I_i (i = 1, 2, \ldots, k)$ which $|p-q|_{A_k} = \sum |p(I_i) - q(I_i)|$. Let $|I_i| = l_i$, $\Delta_i = q(I_i) - p(I_i)$, we can still assume $\frac{\epsilon}{k} \leq l_i \leq \frac{1}{k}$ and $\sum_i |\Delta_i| > \epsilon$ as before. However, the analysis is more complicated than the k-flat case. This is because that the distribution p on the interval I_i is not flat, p on the end of the I_i may mass up. Therefore, we can not use the same analysis as before.

We need a better notion of the partitions instead of the partitions that maximize $\sum |p(I_i) - q(I_i)|$.

First, we restrict $I_i (i = 1, 2, \ldots, k)$ on I_m which is the partition of $[0,1]$ into $k2^m$ equal intervals, $m = \log(1/\epsilon) + O(1)$. If the endpoint of I_i is not on I_m, we can round them to the nearest points on I_m. As shown before, it will introduce at most $\epsilon/100$ error of $|p-q|_{A_k}$.

Now we consider the partition of $[0,1]$ into I_i of length $\leq 1/k$, such that I_i is refinement as above, and the partitions maximize

$$\sum (|\Delta(I_j)|)^2 (k|I_j|)^{-1/20}$$

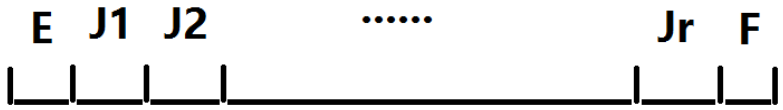
where $\Delta(I_j) = q(I_j) - p(I_j)$.

1. Firstly, we show that the lower bound of $\sum (|\Delta(I_j)|)^2 (k|I_j|)^{-1/20}$ is $O(\epsilon^2/k)$. Using the same analysis as the special case, we can assume that the length of each interval of the optimal partition with respect to $|p-q|_{A_k}$ is $\leq \frac{1}{k}$ Therefore, assume I'_j is the optimal partition, $k|I'_j| \leq 1$, $\sum |\Delta(I'_j)| > \epsilon$, we have

$$\sum (|\Delta(I_j)|)^2 (k|I_j|)^{-1/20} \geq \sum (|\Delta(I'_j)|)^2 (k|I'_j|)^{-1/20} \geq \frac{(\sum |\Delta(I'_j)|)^2}{O(k)} > \frac{\epsilon^2}{O(k)} = O(\epsilon^2/k)$$

2. Secondly, we show that for each interval I_i, there exists m such that $\exists J \in I_m$, the length $|J|$ and the discrepancy $\Delta(J)$ is approximately $|I_i|$ and $\Delta(I_i)$.

   For I_i, we pick m such that $\frac{|I_i|}{2c} \leq \frac{1}{k2^m} < \frac{|I_i|}{c}$ for some appropriate c. Let $J_1, J_2, \ldots, J_r$ be the interval of I_m in I_i, $c \leq r \leq 2c$, the E, F be the interval in the two end of I_i as shown below.



   We know that $|\Delta(I_i)| = |\Delta(E) + \sum \Delta(J_i) + \Delta(F)| \leq |\Delta(E)| + \sum |\Delta(J_i)| + |\Delta(F)|$

   If there exists J_j such that $|\Delta(J_j)| > |\Delta(I_i)|/4c$, we have

$$|\Delta(J_j)|^2 2^{m/20} = |\Delta(J_j)|^2 (k|J_j|)^{-1/20} > (|\Delta(I_i)|/4c)^2 (k\frac{|I_i|}{c})^{-1/20} = O((|\Delta(I_i)|)^2 (k|I_i|)^{-1/20})$$

   Otherwise, if every J_j,$|\Delta(J_j)| \leq |\Delta(I_i)|/4c$, then we have

$$|\Delta(I_i)| \leq |\Delta(E)| + \sum |\Delta(J_i)| + |\Delta(F)| \leq |\Delta(E)| + |\Delta(I_i)|/2 + |\Delta(F)|$$

   which means $|\Delta(E)| > |\Delta(I_i)|/4$ or $|\Delta(F)| > |\Delta(I_i)|/4$.

   Without loss of generality, we assume $|\Delta(E)| > |\Delta(I_i)|/4$. Note that $|E| < \frac{1}{k2^m} < \frac{|I_i|}{c}$, we have

$$(|\Delta(E)|)^2 (k|E|)^{-1/20} > (|\Delta(I_i)|/4)^2 (k|I_i|/c)^{-1/20} = (|\Delta(I_i)|)^2 (k|I_i|)^{-1/20} \times \frac{c^{1/20}}{16}$$

   For c such that $\frac{c^{1/20}}{16} > 1$, we have $(|\Delta(E)|)^2 (k|E|)^{-1/20} > (|\Delta(I_i)|)^2 (k|I_i|)^{-1/20}$, which means we can divide I_i into E and the rest of the part, the sum of $\sum (|\Delta(I_j)|)^2 (k|I_j|)^{-1/20}$ will increase, this is contradiction.

   Therefore, there must exist J_j such that $|\Delta(J_j)| > |\Delta(I_i)|/4c$. It means that for each interval I_i, there exists I_m such that

$$|p^{I_m} - q^{I_m}|^2 2^{m/20} \geq |\Delta(J_j)|^2 2^{m/20} > O((|\Delta(I_i)|)^2 (k|I_i|)^{-1/20})$$

Thus,

$$\sum_m |p^{I_m} - q^{I_m}|^2 2^{m/20} \geq \sum O((|\Delta(I_i)|)^2 (k|I_i|)^{-1/20}) \geq O(\epsilon^2/k)$$

which means there exists m such that $|p^{I_m} - q^{I_m}|_2^2 > O(\frac{\epsilon^2}{k2^{m/3}})$.(Otherwise, if every m, $|p^{I_m} - q^{I_m}|_2^2 < O(\frac{\epsilon^2}{k2^{m/3}})$, then $\sum_m |p^{I_m} - q^{I_m}|_2^2 2^{m/20} < O(\epsilon^2/k) \sum_m 2^{-17m/60} < O(\epsilon^2/k)$)

Thus, if $|p - q|_{A_k} > \epsilon$, we have showed that there exists m such that $|p^{I_m} - q^{I_m}|_2^2 > O(\frac{\epsilon^2}{k2^{m/3}})$. This completes the analysis of the general case.