Lecture 21: Upper bound of A_k closeness testing and Introduction to High Dimensional Structured Distributions

Scirbed by: Hsin Lu

May 19th 2017

1 Last Lecture

In last lecture, we covered the lower bound of sample complexity in A_k closeness testing, which is $O(max\{k^{4/5}/\epsilon^{6/d}, k^{1/2}/\epsilon^2\})$.

Today, we'll go into the upper bound of this problem. Then we are going to the next big topic: High Dimensional Structured Distribution.

2 Upper bound of A_k closeness testing

The idea is to take advantage of our small domain. One way to do about this is that we want to find a way to reduce the domain size. Suppose we are going to reduce to the factor of 2, we let: p' and q' be the distribution on the domain of size $n/2$. Formally, we have:

$$p' = \lceil p/2 \rceil, q' = \lceil q/2 \rceil$$

What we do is like merging pairs of bins together, for example, take bin 1 and 2 and merge into a new bin, take bin 3 and 4 and merge into a new bin...Now the number of bins are half of the original number of bins.

Why is this useful is because the we expect A_k distance between $p\prime$ and $q\prime$ is approximately equal to the A_k distance between p and q:

$$|p' - q'|_{A_k} \approx |p - q|_{A_k}$$

Let's first see a simple case where the partition I is the union of pairs (each interval only contains two number). In such case, these two distances are equal, because intervals in the small domain are correspond to the intervals on the bigger domain.

However, you might actually have the equality. Because it could be that the endpoints of some intervals land in the middle of other intervals. However, we can solve this by rounding the endpoints of the interval to the nearest even number. But this may change the A_k distance by at most $2|p_i - q_i|$ in the i^{th} bin. So, using this, we can have an upper bound of the A_k distance between p and q:

$$|p - q|_{A_k} \leq |p' - q'|_{A_k} + 2|p - q|_{1,k} \tag{1}$$

Here $|p - q|_{1,k}$ denotes the sum of the k largest entries in $|p - q|_1$.

The basic idea of the tester is to test one of the following things: Either $|p - q|_{1,k}$ is large or $|p' - q'|_{A_k}$ is large. We can do this recursively:

We first let the number of initial set of samples $m = min\{k, k^{2/3}/\epsilon^{4/5}\}$ and S to be a set of $Poisson(m)$ samples which we'll use to break things into bins. Then we'll use a L_2 tester to test either $p_s = q_s$ or $|p_s - q_s|_2^2 > \frac{\epsilon^2}{k+|S|}$.

For now we can get the sample complexity: $O(m + \frac{(k+m)^{1/\sqrt{m}}}{\epsilon^2})$ where $1/\sqrt{m}$ is the expected L_2 norm of the distributed distribution on the Poisson m samples.

Using this tool and the previous inequalty, we can have a clean algorithm:

1. Let $t = 0, 1, ..., log(n/k)$
2. Let $p^t = \lceil p/2^t \rceil, q^t = \lceil q/2^t \rceil$
3. Apply inequality (1) for every t

Finally we have:

$$|p - q|_{A_k} \leq |p' - q'|_{A_k} + 2|p^0 - q^0|_{1,k} + ... + 2|p^{t-1} - q^{t-1}|_{1,k} + 2|p^t - q^t|_{1,k}$$

For now we can rewrite our A_k tester in terms of "$1, k$ tester", all we have to do is to distinguish $p = q$ from $|p_i - q_i|_{1,k} > \frac{\epsilon}{4t}$ for some $0 \leq i \leq t$.

Note that the ϵ are smaller by a log factor, so we pick up a polylog over n/k term. Now the upper bound ends up being:

$$O((\frac{k^{2/3}}{\epsilon^{4/3}} + \frac{k^{1/2}}{\epsilon^2})polylog(n/k))$$

3 High-dimensional Structured Distribution

When the dimension grows high, things might be difficult due to the computational intractability. However, we are still able to give computational efficient algorithms on some specific distributions like n-dimensional Gaussian, n-dimensional binomial product and Gaussian Mixtures.

We first look into a easy case: n-dimensional Gaussian distribution: $\mathcal{N}(\mu, \Sigma)$. We assume that the covariance matrix is just the identity matrix for simplicity: $\Sigma = I$.

The first we want to do is to figure out the L_1 distance between two such Gaussian distributions. It is not hard to show that:

$$d_{TV}(\mathcal{N}(\mu, I), \mathcal{N}(\mu', I')) \propto min(1, |\mu - \mu'|_2)$$

The proof is quite straightforward because everything here is rotation invariant. We can rotate such that $\mu - \mu'$ is along with the coordinate axis. This reduce our problem to 1-d case, which is trivial.

Learning

To learn such distribution, we have to return $\mathcal{N}(\hat{\mu}, I)$, where $\hat{\mu}$ is the sample mean of N samples. In fact, $\hat{\mu}$ also follows a Gaussian distribution: $\mathcal{N}(\mu, I/\sqrt{N})$. Therefore, if we want the expected value of $|\mu - \mu'|_2^2 = n/N \ll \epsilon^2$, we just need

$$N > n/\epsilon^2$$

.

Testing

The next thing to do is to testing. Assume we have samples from $N(\mu, I)$, we need to distinguish either $\mu = 0$ or $|\mu|_2 > \epsilon$.

A good way to do this is to consider $|\hat{\mu}|_2^2$, because we have:

$$E(|\hat{\mu}|_2^2) = |\mu|_2^2 + n/N$$

.

In order to figure out what is going on we need to consider the variance. Basically, we have n coordinates independent with each other, so the variance should be like n/N^2. So in order to be able to distinguish the two cases, we

the difference of the mean, which is ϵ^2, to be much bigger than the square root of the variance, which is $\sqrt{n}/N$. Thus, we have that

$$N > \sqrt{n}/\epsilon^2$$

. This is nice because it is smaller than what we need in learning.