0368.416701 Sublinear Time Algorithms November 2, 2015

Lecture 3

Lecturer: Ronitt Rubinfeld *Scribe: I. Ben-Bassat, O. Ben-Eliezer, S. Papini, R. Perlman*

1 Introduction

1.1 Norms

In the previous lesson we saw two norms that are commonly used to measure distance between distributions:

- ℓ_1: $\|p-q\|_1 := \sum |p_i - q_i|$
- ℓ_2: $\|p-q\|_2 := \sqrt{\sum (p_i - q_i)^2}$

It is known that

$$\|p-q\|_2 \leq \|p-q\|_1 \leq \sqrt{n}\|p-q\|_2$$

Today we will prove a result with the ℓ_2-norm. We have seen that ℓ_2 might be a problematic distance measure in some cases, but it does have its uses (in the area of databases, for examples), and moreover results with ℓ_2 sometimes imply results in ℓ_1, using the above relation.

1.2 Background for testing uniformity

Our problem for today is that of testing uniformity. The setting of the problem is the following:

- There is a "black box", called p, that generates independent identically distributed samples.
- The domain D of p is of size n, which is *known*. For simplicity assume that $D = [n] = \{1, \ldots, n\}$.
- The probabilities $p_i = \Pr(p \text{ outputs } i)$ are *unknown*.

The domain can be very large, so we might only want to learn some properties of the distribution p without knowing the values of the p_i.

In the previous lesson we saw a naive algorithm to find the values of the p_i by taking a sample and estimating that p_i is the fraction of appearences of i in the sample. This method gives a good approximation, but requires a large sample: to see all possible elements, we will definitely need a sample of size $\Omega(n)$. In fact, to get an estimate $(q_1, \ldots, q_n)$ of the probabilities, which is of distance less than ϵ from $(p_1, \ldots, p_n)$ in the ℓ_1-norm, the sample size should be $\Omega(n/\epsilon^2)$.

We will see that testing uniformity (that is, testing whether the distribution p is uniform) requires a sample of constant size when working with ℓ_2-norm. With ℓ_1-norm, the required sample size is $\Omega(\sqrt{n})$, but we will use our proof for ℓ_2 to prove for ℓ_1 as well.

Note: since we will only deal with ℓ_2-norm in this lecture, we won't distinguish between $||\cdot||$ and $||\cdot||_2$.

2 The test

Our goal is to find a test whose behavior is as follows:

- If $p = U_{[n]}$ output *PASS* with probability at least 3/4.
- If $\|p - U_{[n]}\|_2^2 \geq \epsilon^2$ output *FAIL* with probability at least 3/4.

Where $U_{[n]}$ is the uniform distribution on $[n]$. Observe that

$$\|p - U_{[n]}\|_2^2 = \sum_{i=1}^{n} (p_i - 1/n)^2 = \sum_i p_i^2 - \sum_i \frac{2}{n} p_i + \sum_i \frac{1}{n^2} = \sum_i p_i^2 - \frac{1}{n} = \|p\|_2^2 - \frac{1}{n}$$

Note that the *collision probability* of p, that is, the probability that two values sampled from p are equal, is also equal to $\|p\|_2^2$. In other words, for any distribution p, the squared ℓ_2-distance between p and the uniform distribution is equal, up to an additive term of $1/n$, to the collision probability of p. In particular, the collision probability of $U_{[n]}$ is $1/n$.

Now we describe an algorithm for testing uniformity. Some of the details will only be filled later.

Algorithm 1 Testing Uniformity

1: $x_1 \ldots, x_s \leftarrow$ *s samples from* P
2: $\hat{c} \leftarrow$ *estimate of* $\|p\|_2^2$ *from* $x_1 \ldots, x_s$
3: **if** $\hat{c} < \frac{1}{n} + \delta$ **then return** PASS
4: **else return** FAIL

There are some gaps that need to be filled:

1. What is the number of samples s?
2. How to estimate?
3. What δ to take?

We will answers those questions in a reverse order.

Finding δ

Assume that after step 2 of the algorithm, $\left|\hat{c} - \|p\|_2^2\right| < \Delta$. Under this assumption:

- If p is the uniform distribution, then $\hat{c} < \frac{1}{n} + \Delta$.
- If p is ϵ-far from uniform (i.e. if $\|p - U_{[n]}\|_2 \geq \epsilon$), then $\hat{c} > \|p\|_2^2 - \Delta = \|p - U_{[n]}\|_2^2 + \frac{1}{n} - \Delta \geq \epsilon^2 + \frac{1}{n} - \Delta$.

We want the algorithm to work correctly when the above assumption holds. This means that we need to have $\frac{1}{n} + \Delta \leq \frac{1}{n} + \delta$ (to handle the case of PASS) and $\frac{1}{n} + \delta \leq \epsilon^2 + \frac{1}{n} - \Delta$ (to handle the case of a FAIL). These two inequalities are true if and only if $\Delta \leq \delta \leq \epsilon^2/2$. Moreover, we certainly want to have Δ as large as possible, since Δ is essentially the "margin of error" that we allow in step 2 of the algorithm. Thus, we take $\Delta = \delta = \epsilon^2/2$, which settles the third question.

Finding the number of samples

We estimate the collision probability by choosing $\hat{c}$ to be the number of collisions in $x_1, \ldots, x_s$. Now, we would like to express $\hat{c}$ as a sum of indicators. For that we will make a few definitions: Let $\sigma_{i,j}$ denote the indicator for the event $x_i = x_j$. That is, we had a collision in places i, j. Let S_k denote the set of all subsets of $\{1, \cdots, s\}$ of size k. Now,

$$\hat{c} = \frac{1}{\binom{s}{2}} \sum_{\{i,j\} \in S_2} \sigma_{i,j}$$

It follows that

$$\begin{aligned}\mathbb{E}[\hat{c}] &= \mathbb{E}\left[\frac{1}{\binom{s}{2}}\sum_{\{i,j\}\in S_2}\sigma_{i,j}\right] \\ &= \frac{1}{\binom{s}{2}}\sum_{\{i,j\}\in S_2}\mathbb{E}\left[\sigma_{i,j}\right]\end{aligned}$$

We know that $\mathbb{E}\left(\sigma_{i,j}\right) = \|p\|^2$, the collision probability. So, we get

$$\begin{aligned}\mathbb{E}[\hat{c}] &= \frac{1}{\binom{s}{2}}\sum_{\{i,j\}\in S_2}\|p\|^2 \\ &= \frac{\binom{s}{2}}{\binom{s}{2}}\|p\|^2 = \|p\|^2\end{aligned}$$

The expected value of our estimate, is indeed $\|p\|^2$, what we wanted to estimate!

We would like to give an upper bound on $\hat{c}$. Since $\{\sigma_{i,j}\}$ are not independent (for example, if $\sigma_{1,10} = 1$ and $\sigma_{10,11} = 1$, then we get that $\sigma_{1,11} = 1$), then we cannot use Chernoff bound. The Markov bound isn't tight enough for our need, so we use Chebyshev bound.

In order to use it, we need to analyze the variance. Recall some known probabilistic quantities:

Definition 2.1 *The variance, $Var(X)$ is defined as $\mathbb{E}\left[X - \mathbb{E}[X]\right]^2 = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2$*

Definition 2.2 *The covariance, $Cov(X,Y)$ is defined as $\mathbb{E}\left[X - \mathbb{E}[X]\right]\left(Y - \mathbb{E}[Y]\right) = \mathbb{E}[XY] - \left(\mathbb{E}[X]\right)\left(\mathbb{E}[Y]\right)$*

Corollary 2.3 *$Var(X)$ and $Cov(X,Y)$ have the following properties:*

- $Var(\lambda X) = \lambda^2 Var(X)$
- $Cov(X,X) = Var(x)$
- $Cov(X, \lambda Y + Z) = \lambda Cov(X,Y) + Cov(X,Z)$
- *If X, Y are independent then $Cov(X,Y) = 0$.*

Now, we can go on and calculate the variance of $\hat{c}$:

$$\begin{aligned}Var(\hat{c}) &= Var\left(\frac{1}{\binom{s}{2}}\sum_{\{i,j\}\in S_2}\sigma_{i,j}\right) \\ &= \frac{1}{\binom{s}{2}^2}Var\left(\sum_{\{i,j\}\in S_2}\sigma_{i,j}\right) \\ &= \frac{1}{\binom{s}{2}^2}Cov\left(\sum_{\{i,j\}\in S_2}\sigma_{i,j}, \sum_{\{k,l\}\in S_2}\sigma_{k,l}\right) \\ &= \frac{1}{\binom{s}{2}^2}\sum_{\{i,j\}\in S_2,\{k,l\}\in S_2}Cov\left(\sigma_{i,j}, \sigma_{k,l}\right)\end{aligned}$$

Now look at $\{i,j\}, \{k,l\}$. We identify 3 cases:

1. $|\{i,j,k,l\}| = 4$. Which means $\{i,j\},\{k,l\}$ are distinct.
2. $|\{i,j,k,l\}| = 3$. Which means $\{i,j\},\{k,l\}$ have one common items.
3. $|\{i,j,k,l\}| = 2$. Which means $\{i,j\} = \{i,j\}$.

We will split the sum above into these 3 cases.

Case 1

$$\frac{1}{\binom{s}{2}^2} \sum_{\{i,j,k,l\} \in S_4} Cov\left(\sigma_{i,j}, \sigma_{k,l}\right) \quad = \quad 0$$

This equality comes from the fact that $\sigma_{i,j}, \sigma_{k,l}$ are independent random variables, because the samples are independent.

Case 2

How many times does $\sigma_{i,j}\sigma_{j,k}$ appear with $\{i,j,k\}$ distinct? Basic combinatorics: s possibilities to choose the repeated index j, and then $(s-1)(s-2)$ to choose i,k (order is important). So terms of this case appear $s(s-1)(s-2)$ times overall. We would like to calculate the covariance $Cov(\sigma_{i,j}, \sigma_{j,k})$.

$$Cov(\sigma_{i,j}, \sigma_{j,k}) = \mathbb{E}[\sigma_{i,j}\sigma_{j,k}] - \left(\mathbb{E}[\sigma_{i,j}]\right)\left(\mathbb{E}[\sigma_{j,k}]\right)$$

Notice that $\sigma_{i,j}\sigma_{j,k}$ is just the indicator for the event $x_i = x_j = x_k$, which happens with probability $\sum_\alpha p_\alpha^3$, which we will denote simply as d. $\mathbb{E}[\sigma_{i,j}] = ||p||^2$, as we've seen before. Putting all together, the contribution of this case is:

$$\frac{1}{\binom{s}{2}^2} s(s-1)(s-2)\left(d - \|p\|^4\right)$$

Case 3

The contribution in this case is:

$$\begin{aligned} & \frac{1}{\binom{s}{2}^2} \sum_{\{i,j\} \in S_2} Cov\left(\sigma_{i,j}, \sigma_{i,j}\right) \\ = \quad & \frac{1}{\binom{s}{2}^2} \sum_{\{i,j\} \in S_2} Var\left(\sigma_{i,j}\right) \\ = \quad & \frac{1}{\binom{s}{2}^2} \binom{s}{2} \|p\|^2(1 - \|p\|^2) \end{aligned}$$

The last equality holds since the variance of an indicator with probability q is just $q(1-q)$. In our case, the probability is $\|p\|^2$ and so the variance is $\|p\|^2(1-\|p\|^2)$.

Total variance

The calculated variance from all the cases is:

$$Var(\hat{c}) = \frac{1}{\binom{s}{2}^2} \binom{s}{2} \|p\|^2(1 - \|p\|^2) + \frac{1}{\binom{s}{2}^2} s(s-1)(s-2)\left(d - \|p\|^4\right)$$

We bound this from above for $s > 1$:

$$Var(\hat{c}) \leq cs^{-1}$$

where c is a constant (not dependent on p).

Chebyshev and choosing s

Chebyshev's inequality gives:

$$Pr\left[\left|\hat{c} - \frac{1}{n}\right| \geq k\sqrt{Var(\hat{c})}\right] < \frac{1}{k^2}$$

We would like to get a probability bound of say $\frac{1}{4}$, so we will take $k = 2$. If we choose s such that $k\sqrt{Var(\hat{c})} < \frac{\epsilon^2}{2}$, we will get that with probability at least $3/4$ we have $\left|\hat{c} - \frac{1}{n}\right| < k\sqrt{Var(\hat{c})} < \frac{\epsilon^2}{2}$ which is exactly what we wanted!

So, the requirement is $2\sqrt{Var(\hat{c})} < \frac{\epsilon^2}{2}$, or $Var(\hat{c}) < \frac{\epsilon^4}{16}$. Since we know $Var(\hat{c}) \leq cs^{-1}$ if we take $s = \frac{16c}{\epsilon^4}$, we get a tester with the specified requirements, using $O(\frac{1}{\epsilon^4})$ samples.
In recent results, the upper bound was reduced to $O(\frac{1}{\epsilon^2})$, using a more careful analysis.