COMS E6998-9: Algorithmic Techniques for Massive Data Nov 5, 2015

Lecture 17 — Sublinear Time

Instructor: *Alex Andoni* Scribes: *Michael J Curry*

1 Introduction

At the beginning of the lecture, we briefly went over some upper and lower bounds on distortion for embeddings of various distances into l_1. Then, we discussed the setting for sublinear time algorithms, one particular problem (distribution testing), and specific algorithms for distinguishing between uniform and sufficiently non-uniform distributions.

2 Sublinear Time Algorithms

We've been concerned with situations where the size of the input is much too large to deal with using conventional algorithms. We've considered streaming algorithms, where each part of the input is seen once and the goal is to approximate an answer given space constraints. Now we'll turn to truly sublinear algorithms — only looking at a subset of the input. This might be necessary due to resource constraints (like on a router where even simple operations take unacceptably long). The data itself might also come as a sample, as in natural experiments or in the setting for machine learning where the training data is assumed to be drawn from some unknown distribution.

Broadly, there are two types of sublinear algorithms: the "classic" type, which examine a subset of the data and return an approximate output; and "property testing", in which the goal is to verify whether some object has a certain property. We'll consider the problem of distribution testing.

3 Testing for Uniformity

It's hard to precisely tell whether a distribution is uniform, but we can accept some approximation and try to distinguish between uniform and "sufficiently non-uniform" distributions.

3.1 Total Variation

Suppose we treat (discrete) distributions over $[n]$ as vectors of probabilities. Then we consider a distribution D "sufficiently non-uniform" if $\|D-U_n\|_1 \geq \epsilon$. The l_1 distance here is a proxy for the Total Variation distance. Suppose we define our test as a subset $T \subset [n]$; if $x \in T$ when sampling from one distribution but not the other then we can distinguish the distributions. Then the total variation distance is defined as

$$TV(A, B) = \max_{T \subset [n]} |Pr_A[x \in T] - Pr_B[x \in T]| \tag{1}$$

Claim 1. $TV(A, B) = \frac{1}{2}\|A - B\|_1$

3.2 First Attempt at an Algorithm

We could estimate the distribution D empirically, and then compute the distance $\|\hat{D} - U\|_1$. This is not a very good algorithm. We need at least $n/2$ samples; otherwise at least half the coordinates are guaranteed to be zero, resulting in an estimate that is far from uniform. The χ^2 test in classical statistics also requires $\Omega(n)$ samples.

3.3 Second Attempt

Algorithm 1 UNIFORM

Require: $n, m, x_1 \ldots x_n$
  $C \leftarrow 0$
  for $i = 0$ to m **do**
    for $j = 0$ to m **do**
      if $x_i = x_j$ **then**
        $C \leftarrow C + 1$
      end if
    end for
  end for
  if $C < \frac{am^2}{n}$ **then**
    return uniform
  else
    return nonuniform
  end if

We can actually estimate uniformity using only $O_\epsilon(\sqrt{n})$ samples. The idea is to sample and count the number of collisions: a nonuniform distribution will have more collisions. The amount of sampling required is connected to the famous "birthday paradox" — in a uniform distribution, we would expect collisions to start appearing at around $\sqrt{n}$ samples.

3.3.1 Analysis

First, think about the l_2 distance between distributions.

Claim 2. *If $D = U_n$ then $\|D - U_n\|_2 = 0$. But if $\|D - U_n\|_1 \geq \epsilon$, then $\|D - U_n\|_2 > \frac{\epsilon^2}{n}$.*

Proof. The first part is obvious. For the second part, note that

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \leq \left(\sum_{i=1}^{n} 1\right)^{1/2} \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$$

So $\forall x : \|x\|_2 \geq \frac{\|x\|_1}{\sqrt{n}}$. □

Claim 3. $\|D - U_n\|_2^2 = \|D\|_2^2 - \frac{1}{n}$

Proof. (The terms of D are denoted P_i, because they represent probabilities.)

$$\begin{aligned}\|D - U_n\|_2^2 &= \|D\|_2^2 + \|U_n\|_2^2 - 2D \cdot U_n \\ &= \|D\|_2^2 + \sum_{k=1}^{n} \left(\frac{1}{n}\right)^2 - 2\sum_{k=1}^{n} \frac{P_k}{n} \\ &= \|D\|_2^2 + \frac{1}{n} - \frac{2}{n}\sum_{k=1}^{n} P_k \\ &= \|D\|_2^2 - \frac{1}{n}\end{aligned}$$

□

The upshot is that $\|D_2\|_2^2 = \frac{1}{n}$ when uniform and $\|D_2\|_2^2 > \frac{1}{n} + \frac{\epsilon^2}{n}$ when non-uniform, so we just need to be able to distinguish these cases.

Lemma 4. $\frac{1}{M} \times C$ *allows us to distiguish between the two cases above, as long as* $m = \Omega(\frac{\sqrt{n}}{\epsilon^4})$. *$M$ is the normalization constant* $M = \binom{m}{2}$.

Proof. As before, we first show that the estimator is unbiased and then bound its variance. Define σ_{ij} to be the indicator variable of the event $x_i = x_j$. Also $Z = \frac{1}{M}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\sigma_{ij}$.

$$\begin{aligned}M \cdot E[Z] &= E\left[\sum_{i=1}^{n}\sum_{j=i+1}^{n}\sigma_{ij}\right] \\ &= \sum_{i=1}^{n}\sum_{j=i+1}^{n}[\sigma_{ij}] \\ &= \sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{k=1}^{n} P_k P_k \\ &= \sum_{i=1}^{n}\sum_{j=i+1}^{n}\|D\|_2^2 \\ &= \binom{m}{2}\|D\|_2^2\end{aligned}$$

The variance is a little trickier to bound than in the past. Essentially, we'll break up a sum into 3 terms and bound each of them.

$$\begin{aligned}E\left[Z^2\right] &= \frac{1}{M^2}E\left[\sum_{i_1<j_1}\sum_{i_2<j_2}\sigma_{i_1 j_1}\sigma_{i_2 j_2}\right] \\ &= \frac{1}{M^2}E\left[\sum_{i_1=i_2<j_1=j_2}\sigma_{i_1 j_1}^2 + \sum_{|\{i_1,i_2,j_1,j_2\}|=3}\sigma_{i_1 j_1}\sigma_{i_2 j_2} + \sum_{\{i_1\}\cap\{i_2\}\cap\{j_1\}\cap\{j_2\}=\emptyset}\sigma_{i_1 j_1}\sigma_{i_2 j_2}\right]\end{aligned}$$

Bringing the expectation operator inside, we can bound the first term:

$$E\left[\sum_{i<j}\sigma_{ij}^2\right] = M\|D\|_2^2$$

The second term:

$$\begin{aligned}2E\left[\sum_i\sum_{i<j_1\neq j_2}\sigma_{ij_1}\sigma_{ij_2}\right] &= 2\sum_i\sum_{i<j_1\neq j_2}\sum_k P_kP_kP_k\\ &\leq 2m^2\|D\|_3^3 \leq 2m^3\left(\|D\|_2^2\right)^{3/2}\end{aligned}$$

In the third term, everything in the summand is independent, so

$$E\left[\sum_{\{i_1\}\cap\{i_2\}\cap\{j_1\}\cap\{j_2\}=\emptyset}\sigma_{i_1j_1}\sigma_{i_2j_2}\right] \leq M^2\left(\|D\|_2^2\right)$$

For convenience let $d = \|D\|_2^2$. The variance of Z is at most

$$\frac{1}{M^2}\left(Md + 2m^3d^{3/2} + M^2d^2\right) - d^2 \leq \frac{d}{M} + \frac{8d^{3/2}}{m} \leq \frac{d}{M} + \frac{8d^2}{md^{1/2}} \leq \frac{d}{M} + \frac{8d^2\sqrt{n}}{m}$$

The lecture ended here. $\square$