6.5240 Sublinear Time Algorithms October 24, 2022

Lecture 12

Lecturer: Ronitt Rubinfeld *Scribe: Kenneth Moon*

Starting with this lecture, we will begin a new topic: testing properties of distributions. We often want to know many different properties about distributions from which we can only realistically see a small sample. For instance, we might want to know how many butterfly species exist in the world (the size of the support of a distribution), or whether some lottery numbers show up more often than others (distribution uniformity). In this lecture, we will focus on testing uniformity.

1 Definition of Testers on Distributions

We have previously seen testers on discrete objects such as graphs or lists. In those cases, we specified the criteria for testers by ensuring (with some constant probability) that the tester will reject inputs that look very different from inputs which satisfy the given property. Similarly, for testers on distributions, we need some way to quantify how different two distributions are, and reject distributions that are very different from distributions satisfying a given property. There are many measures of "different-ness" but in this lecture we will focus on the L_1 and L_2 metric. They can be defined in a generalized way as follows:

Definition 1 *For any p, the L_p metric between two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ over domain $\mathcal{D}$ is:*

$$||\mathcal{P}_1 - \mathcal{P}_2||_n = \left(\sum_{i \in \mathcal{D}} |\Pr_{P \sim \mathcal{P}_1}[P = i] - \Pr_{Q \sim \mathcal{P}_2}[Q = i]|^p \right)^{\frac{1}{p}}$$

Now, it is important to address some nuances of the L_1 and L_2 metrics. Although they are defined similarly, their values can be very different for the same pair of distributions in unintuitive ways. For example, consider two distributions on $[2n]$: one is evenly distributed on numbers from 1 to n, while the other is evenly distributed on numbers from $n+1$ to $2n$. These distributions seem to be very different: they never even overlap. The L_1 metric seems to confirm this, giving a distance of 2 between the distributions. However, the L_2 metric gives a result of $\frac{2}{\sqrt{n}}$, vanishingly small.

With these metrics, we can formally define a tester of some property of distributions as follows.

Definition 2 *A tester for property Π is an algorithm which takes independent random samples from input distribution $\mathcal{P}$ and outputs a decision to accept or reject it.*

The tester is parameterized by a distance function of distributions dist *and a threshold* ϵ*. It must have the following guarantees:*

- *If* $\mathcal{P} \in \Pi$*, it must accept with probability at least* $\frac{2}{3}$*.*
- *If* $\min_{\mathcal{P}' \in \Pi} \text{dist}(\mathcal{P}, \mathcal{P}') > \epsilon$*, it must reject with probability at least* $\frac{2}{3}$*.*

The complexity of the tester is defined as the number of i.i.d. samples which must be drawn from an input distribution before the tester makes a decision.

2 Lower Bounds for Testing Uniformity in L_1

Let us now focus on testing uniformity for distributions over a support of size n. By our previous definition, we want an algorithm that takes some i.i.d. samples from an input distribution and outputs a decision such that uniform distributions are accepted at least $\frac{2}{3}$ of the time, while distributions that are ϵ-far from the uniform distribution are rejected at least $\frac{2}{3}$ of the time.

2.1 Naive Algorithm Bound

We will begin by proving a lower bound for the following naive algorithm which relies on computing the empirical distribution given m samples over a support S. The algorithm is a tester in the L_1 metric.

Algorithm 1 Plug-in algorithm

Initialize $count(x)$ for all $x \in S$ to 0.
for $i \in [m]$ **do**
  Sample $X_i \sim \mathcal{P}$
  Increment $count(X_i)$
end for
Let $\mathcal{P}_e$ be the distribution such that $\Pr_{P \sim \mathcal{P}_e}[P = x] = \frac{count(x)}{m}$.
Reject if $||\mathcal{P}_e - \mathcal{U}_{\mathcal{D}}||_1 > \epsilon$; accept otherwise

Now, we shall prove the following theorem:

Theorem 3 $E[||\mathcal{P}_e - \mathcal{P}||_1] \leq \sqrt{\frac{n}{m}}$

Proof Let $p(x) = \Pr_{P \sim \mathcal{P}}[P = x]$, the probability we draw x from our input distribution.

$$E[||\mathcal{P}_e - \mathcal{P}||_1] = \sum_{x \in S} E\left[\left|\frac{count(x)}{m} - p(x)\right|\right]$$

By Jensen's inequality:

$$E\left[\left|\frac{count(x)}{m} - p(x)\right|\right] \leq \sqrt{E\left[\left(\frac{count(x)}{m} - p(x)\right)^2\right]}$$

Furthermore, the expected value of $\frac{count(x)}{m}$ is $p(x)$, thus:

$$E\left[\left(\frac{count(x)}{m} - p(x)\right)^2\right] = \text{Var}\left(\frac{count(x)}{m}\right)$$

$count(x)$ is the sum of independent Bernoulli variables with parameter $p(x)$, which means its variance is $mp(x)(1-p(x))$. So:

$$\begin{aligned} E[||\mathcal{P}_e - \mathcal{P}||_1] &\leq \sum_{x \in S} \sqrt{\frac{p(x)(1-p(x))}{m}} \\ &\leq \sum_{x \in S} \sqrt{\frac{p(x)}{m}} \end{aligned}$$

By the Cauchy-Schwarz inequality:

$$\begin{aligned} \left(\sum_{x \in S} \sqrt{\frac{p(x)}{m}}\right)^2 &\leq n \cdot \sum_{x \in S} \left(\sqrt{\frac{p(x)}{m}}\right)^2 \\ &\leq \frac{n \cdot \sum_{x \in S} p(x)}{m} \\ &\leq \frac{n}{m} \end{aligned}$$

Thus:

$$\begin{aligned} \sum_{x \in S} \sqrt{\frac{p(x)}{m}} &\leq \sqrt{\frac{n}{m}} \\ E[||\mathcal{P}_e - \mathcal{P}||_1] &\leq \sqrt{\frac{n}{m}} \end{aligned}$$

■

Now, note that this is terrible. If $m \in o\left(\frac{n}{\epsilon^2}\right)$, the expected value of our deviation from the true distribution is $\omega(\epsilon)$. This suggests that it is reasonable to expect that even a distribution which is ϵ-far from uniform can be confused with a uniform distribution. Thus, this naive algorithm must take at least $\Omega\left(\frac{n}{\epsilon^2}\right)$ samples.

2.2 General Lower Bound

We will not derive a rigorous proof for the general lower bound for testing uniformity in L_1, but we will give some intuition. Note that until the tester witnesses an element in the support more than once, the tester cannot infer anything meaningful about the probability of each element appearing, at least among the elements that have appeared so far. This is due to the fact that all the labels could be permuted, yet the distance of the underlying distribution from uniform would be unchanged.

Now, recall the birthday paradox: in a uniform distribution, it takes $\Omega(\sqrt{n})$ samples for a collision to happen with $\Omega(1)$ probability. It turns out that it is possible to construct a distribution by slightly perturbing the uniform distribution to become ϵ-far from uniform in L_1, yet still maintain this bound. Thus, $\Omega(\sqrt{n})$ samples are needed for any tester in L_1.

3 Uniformity Tester in L_2

Let us now turn our attention to the L_2 metric. First, note that the L_2 metric is intrinsically related to the collision probability; that is, the probability that we draw two independent samples which have the same value. This can be used to bound the collision probability of distributions that are ϵ-far from uniform:

Theorem 4 *If distribution $\mathcal{P}$ has domain $[n]$ and is ϵ-far from uniform in L_2, then it has a collision probability of at least $\epsilon^2 + \frac{1}{n}$.*

Proof Let $p(i)$ be the probability that $X = i$ if $X \sim \mathcal{P}$. Then:

$$\begin{aligned} ||\mathcal{P} - \mathcal{U}_{[n]}||_2^2 &= \sum_{i=1}^{n} \left(p(i) - \frac{1}{n} \right)^2 \\ &= \sum_{i=1}^{n} p(i)^2 - \frac{2}{n} \sum_{i=1}^{n} p(i) + \sum_{i=1}^{n} \frac{1}{n} \end{aligned}$$

Note that $\sum_{i=1}^{n} p(i) = 1$. Thus:

$$||\mathcal{P} - \mathcal{U}_{[n]}||_2^2 = \left(\sum_{i=1}^{n} p(i)^2 \right) - \frac{1}{n}$$

Now, using ϵ-far's definition:

$$\begin{aligned} ||\mathcal{P} - \mathcal{U}_{[n]}||_2^2 &> \epsilon^2 \\ \sum_{i=1}^{n} p(i)^2 &> \epsilon^2 + \frac{1}{n} \end{aligned}$$

Note that $\sum_{i=1}^{n} p(i)^2$ is equivalent to the collision probability of $\mathcal{P}$. ■

This naturally leads to the following idea: what if we tried to estimate the collision probability to indirectly estimate how far a distribution is from uniform?

To estimate collision probability, we can count how many colliding pairs there are in a given sample of size s, and average it over the total number of pairs which is $\binom{s}{2}$. We will call this our empirical collision rate or $\hat{c}$ (whereas the true collision rate is c). We then set a threshold: if $\hat{c} > \frac{1}{n} + \Delta$ for some Δ, we reject the distribution.

3.1 Analysis

Our tester must balance between two opposing demands: accepting uniform distribution, while rejecting ϵ-far ones. To accept uniform distributions, we must guarantee with some constant probability that $\hat{c} \leq \frac{1}{n} + \Delta$. Since $\frac{1}{n}$ is the collision probability of uniform distributions, this is the same as demanding that our estimator is within Δ of the true collision probability.

Now, to reject ϵ-far distributions, we must have $\hat{c} > \frac{1}{n} + \Delta$. Since the colliding probability of our distribution must be more than $\frac{1}{n} + \epsilon^2$ by Theorem 4, this is the same as demanding that our estimator is within $\epsilon^2 - \Delta$ of the true collision probability. To balance these demands, it is clear we must set Δ to $\frac{\epsilon^2}{2}$.

However, how do we get such a guarantee on our estimator? We can't use Chernoff bounds, since the pairs we choose are not independent. Let us use Chebyshev bounds instead. To use this bound, we need to find $\mathrm{Var}(\hat{c})$. Let $\sigma_{i,j}$ be the indicator variable for the event that the i^{th} and j^{th} sample collide. Then, $\hat{c} = \frac{1}{\binom{s}{2}} \sum_{i,j} \sigma_{i,j}$. So:

$$\mathrm{Var}(\hat{c}) = \frac{1}{\binom{s}{2}^2} \mathrm{Var}\left(\sum_{i,j} \sigma_{i,j} \right)$$

Let $\bar{\sigma}_{i,j} = \sigma_{i,j} - E[\sigma_{i,j}]$. Then, by the definition of variance:

$$\mathrm{Var}\left(\sum_{i,j} \sigma_{i,j} \right) = E\left[\left(\sum_{i,j} \bar{\sigma}_{i,j} \right)^2 \right]$$

By linearity of expectation, we can expand expression to cover cases where pairs overlap completely, partially overlap, and are independent:

$$= E\left[\sum_{i,j} \bar{\sigma}_{i,j}^2 \right] + E\left[\sum_{i,j,k,l} \bar{\sigma}_{i,j} \bar{\sigma}_{k,l} \right] + E\left[\sum_{i,j,k} \bar{\sigma}_{i,j} \bar{\sigma}_{j,k} \right]$$

Note that the indices in each sum are unique. For example, in the second sum, i, j, k and l are all different numbers.

This may look scary because there are $O(n^4)$ terms. However, we will show that most of them don't matter. Let us bound the first sum:

$$\begin{aligned} E\left[\sum_{i,j} \bar{\sigma}_{i,j}^2\right] &\leq E\left[\sum_{i,j} \sigma_{i,j}^2\right] \\ &\leq \sum_{i,j} E\left[\sigma_{i,j}^2\right] \\ &\leq \sum_{i,j} E\left[\sigma_{i,j}\right] \\ &\leq \binom{s}{2} \cdot c \end{aligned}$$

Now, let us take a look at the second sum. Since $\bar{\sigma}_{i,j}$ and $\bar{\sigma}_{k,l}$ are independent (i, j, k, l are all unique), $E[\bar{\sigma}_{i,j}\bar{\sigma}_{k,l}] = E[\bar{\sigma}_{i,j}]E[\bar{\sigma}_{k,l}]$, which is 0 since the expectation of all $\bar{\sigma}$ is 0. This is promising since we just wiped out most of the terms in the sum! We will show a bound on the third sum next lecture, which will give us the tester's complexity.