CSE 522: Sublinear (and Streaming) Algorithms **Spring 2014**

Lecture 7: Testing Distributions

April 21, 2014

Lecturer: Paul Beame *Scribe: Paul Beame*

1 Testing Uniformity of Distributions

We return today to property testing and a surprising application of F_2 estimation (or equivalently ℓ_2-norm approximation) for the problem of testing the closeness to uniform of a probability distribution.

We consider discrete distributions on the domain $[n]$. Each such distribution is given by a probability vector

$$p = (p_1, \ldots, p_n).$$

The uniform distribution U on $[n]$ has $U_i = 1/n$ for all $i \in [n]$. The property testing problem we focus on first is to do the following:

- Accept if p is uniform
- Reject if p is ε-far from uniform.

This definition is ambiguous without a notion of distance between distributions.

The most natural notion of distance between distributions is the ℓ_1 distance between their probability vectors, where

$$||p - q||_1 = \sum_i |p_i - q_i| = 2 \cdot \max A \subset [n] p(A) - q(A).$$

$||p-1||_1$ is called the *total variation distance* between distributions, while $\max A \subset [n] p(A) - q(A)$ is called the *statistical distance*. Another notion that we will consider is the ℓ_2 distance between distributions,

$$||p - q||_2 = \sqrt{\sum_i |p_i - q_i|}.$$

While ℓ_1 distance is the most natural, it will turn out that analyzing the ℓ_2 distance will be the most useful algorithmically.

These can be closely related — if $p = (1, 0, \ldots, 0)$ then $||p - U||_1 = 2 - 2/n$ and $||p - U||_2 \approx 1$ — but if $p = (2/n, \ldots, 2/n, 0, \ldots, 0)$ and $q = (0, \ldots, 0, 2/n, \ldots, 2/n)$ then $||p - q||_1 = 2$ but $||p - q||_2$ is only $2/\sqrt{n}$. The latter is the largest gap possible. Observe that for any probability distribution p on $[n]$, $1/\sqrt{n} \leq ||p||_2 \leq ||p||_1 = 1$ and

$$||p - q||_2/n^{1/2} \leq ||p - q||_2 \leq ||p - q||_1 \leq n^{1/2}||p - q||_2.$$

Naive algorithm for ℓ_1 distance from uniformity:
Choose s samples and compute the sample distribution $\tilde{p}$ as an approximation to p.
Compute $||U - \tilde{p}||_1$ to estimate $||p - U||_1$.

The problem with this approach is that $||p - \tilde{p}||_1$ is huge unless $s = \Omega(n)$. (Otherwise, the algorithm is trying to estimate n different quantities with $o(n)$ samples. In this case, every $\tilde{p}$ would be distance $1 - o(1)$ from uniform.)

Alternative idea: Use an algorithm for property testing with respect to ℓ_2 distance to derive an algorithm for property testing with respect to ℓ_2 distance.

The reason for the utility of the ℓ_2 distance is closely related to the connection of F_2 to the sizes of self-joins. Observe that the collision probability for independent samples from p,

$$\mathbb{P}_{a,b\sim p}[a = b] = \sum_{i=1}^{n} p_i^2 = ||p||_2^2.$$

Also $\mathbb{P}_{a,b\sim U}[a = b] = 1/n$. Then the square of the ℓ_2 distance,

$$\begin{aligned}
||p - U||_2^2 &= \sum_{i=1}^{n}(p_i - \frac{1}{n})^2 \\
&= \sum_{i=1}^{n}(p_i^2 - 2\frac{p_i}{n} + \frac{1}{n^2}) \\
&= \sum_{i=1}^{n} p_i^2 - 2\frac{\sum_{i=1}^{n} p_i}{n} + \frac{n}{n^2}) \\
&= ||p||_2^2 - \frac{2}{n} + \frac{1}{n}) \\
&= ||p||_2^2 - \frac{1}{n},
\end{aligned}$$

or equivalently $||p||_2^2 = \frac{1}{n} + ||p - U||_2^2$.

This suggests sampling to get a good estimate of the collision probability, $||p||_2^2$. What error will we need?

ℓ_2 **distance:** If $||p - U||_2 > \varepsilon$ then

$$||p||_2^2 = 1/n + ||p - U||_2^2 > 1/n + \varepsilon^2.$$

On the other hand $||U||_2^2 = 1/n$. In order to separate p from U, we would need to separate $1/n$ from something $> 1/n + \varepsilon^2$, even with error. It suffices to have an additive error in computing $||p||_2^2$ of at most $\varepsilon^2/2$.

ℓ_1 **distance:** If $||p - U||_1 > \varepsilon$ then $||p - U||_2 > \varepsilon/\sqrt{n}$ so $||p - U|_2^2 > \varepsilon^2/n$. Therefore,

$$||p||_2^2 = 1/n + ||p - U||_2^2 > 1/n + \varepsilon^2/n = 1/n(1 + \varepsilon^2).$$

In this case we need to distinguish $1/n$ from something larger than $1/n\ (1 + \varepsilon^2)$. It is natural to consider multiplicative error in this case. With multiplicative error $1 \pm \varepsilon^2/3$ observe that if $||p - U||_1^1 > \varepsilon$ then $||p||_2^2$ would evaluate to strictly more than

$$1/n\ (1 + \varepsilon^2)(1 - \varepsilon^2/3) = 1/n(1 + 2\varepsilon^2/3 - \varepsilon^4/3) \geq 1/n(1 + \varepsilon^2/3)$$

which is the largest that the value could be if $p = U$.

Property testing algorithm for uniformity:
Choose s independent samples $x_1, \ldots, x_s$ from p.
Let $Y_{ij} = \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{otherwise.} \end{cases}$
Output $X = \sum_{i<j} Y_{ij} / \binom{s}{2}$.

Analysis By definition,

$$\mathbb{E}(Y_{ij} = ||p||_2^2.$$

Therefore, since X is the average of the Y_{ij},

$$\mathbb{E}(X) = ||p||_2^2$$

and so X is an unbiased estimator for $||p||_2^2$.

$$\begin{aligned} Var(\binom{s}{2} X) &= Var(\sum_{i<j} Y_{ij}) \\ &= \mathbb{E}([\sum_{i<j} Y_{ij} - \mathbb{E}(\sum_{i<j} Y_{ij})]^2) \\ &= \mathbb{E}([\sum_{i<j} Y_{ij} - \mathbb{E}(Y_{ij})]^2). \end{aligned}$$

We write $\hat{Y}_{ij} = Y_{ij} - \mathbb{E}(Y_{ij}) = Y_{ij} - ||p||_2^2$ and note that $\mathbb{E}(\hat{Y}_{ij}) = 0$. Then

$$\begin{aligned} Var(\binom{s}{2}X) &= \mathbb{E}([\sum_{i<j}\hat{Y}_{ij}]^2) \\ &= \mathbb{E}(\sum_{i<j}\hat{Y}_{ij}\sum_{k<\ell}\hat{Y}_{k\ell}) \\ &= \mathbb{E}(\sum_{i<j}\sum_{k<\ell}\hat{Y}_{ij}\hat{Y}_{k\ell}) \\ &= \sum_{i<j}\mathbb{E}(\hat{Y}_{ij}^2) + \sum_{\substack{i<j \\ k<\ell \\ |\{i,j,k,\ell\}|=3}}\mathbb{E}(\hat{Y}_{ij}\hat{Y}_{k\ell})) + \sum_{\substack{i<j \\ k<\ell \\ |\{i,j,k,\ell\}|=4}}\mathbb{E}(\hat{Y}_{ij}\hat{Y}_{k\ell}). \end{aligned}$$

Now if i, j, k, ℓ are all distinct, the random variables $\hat{Y}_{ij}$ and $|\hat{Y}_{k\ell}$ are independent so $\mathbb{E}(\hat{Y}_{ij}\hat{Y}_{k\ell}) = \mathbb{E}(\hat{Y}_{ij}) \cdot \mathbb{E}(\hat{Y}_{k\ell}) = 0 \cdot 0 = 0$ and hence the third term in the sum is 0.

More generally, for any $i < j$ and $k < \ell$,

$$\begin{aligned} \mathbb{E}(\hat{Y}_{ij}\hat{Y}_{k\ell}) &= \mathbb{E}((Y_{ij} - ||p||_2^2)(Y_{k\ell} - ||p||_2^2)) \\ &= \mathbb{E}(Y_{ij}Y_{k\ell} - ||p||_2^2(\mathbb{E}(Y_{ij} + \mathbb{E}(Y_{k\ell}) + ||p||_2^4 \\ &= \mathbb{E}(Y_{ij}Y_{k\ell} - 2||p||_2^4 + ||p||_2^4 \\ &= \mathbb{E}(Y_{ij}Y_{k\ell} - ||p||_2^4 \\ &< \mathbb{E}(Y_{ij}Y_{k\ell}). \end{aligned}$$

In particular, $\mathbb{E}(\hat{Y}_{ij}^2) \leq \mathbb{E}(Y_{ij}^2) = \mathbb{E}(Y_{ij}) = ||p||_2^2$ since Y_{ij} is an indicator variable. Therefore the first term in the sum is $\binom{s}{2}||p||_2^2$.

Also, for every $i < j$ and $k < \ell$ such that $|\{i, j, k, \ell\}| = 3$, the event $Y_{i,j}Y_{k,\ell}$ is the event that all the samples indexed by them produce the same value. For any three samples, this probability is precisely $\sum_{i=1}^n p_i^3 = ||p||_3^3$. For each of the $\binom{s}{3}$ choices of three samples, there are 6 ways that this can correspond to $i < j$ and $k < \ell$: if $i = \ell$ or $k = j$ then there is only one way to extend this to three samples, if either $i = k$ or $j = \ell$ there are a further two ways to order the remaining indices.

Putting this together, we have

$$Var(\sum_{i<j}Y_{ij}) \leq \binom{s}{2}||p||_2^2 + 6\binom{s}{3}||p||_3^3.$$

Therefore,

$$\begin{aligned} Var(X) &= Var(\sum_{i<j} Y_{ij}) / \binom{s}{2}^2 \\ &\leq \frac{\binom{s}{2}||p||_2^2 + 6\binom{s}{3}||p||_3^3}{\binom{s}{2}^2} \\ &= \frac{2||p||_2^2}{s(s-1)} + \frac{4(s-2)||p||_3^3}{s(s-1)} \\ &< \frac{2||p||_2^2}{s(s-1)} + \frac{4||p||_3^3}{s}. \end{aligned}$$

ℓ_2 testing quality approximation for $||p||_2$: This requires an $\varepsilon^2/2$ additive approximation. Observe that $||p||_2^2, ||p||_3^3 \leq 1$ so $Var(X) \leq 1/binoms2 + 4/s < 5/s$ for $s \geq 5$. Therefore, by Chebyshev's inequality,

$$\mathbb{P}[|X - ||p||_2^2| \geq \varepsilon^2/2] \leq \frac{Var(X)}{\varepsilon^2/2)^2} \leq 1/3,$$

for $s = O(\varepsilon^{-2})$, in particular $s = 60/\varepsilon^2$. Therefore, for constant ε, only a constant number of samples are required to test the proximity p to uniform distribution using the ℓ_2 error measure.

ℓ_1 testing quality approximation for $||p||_2$: This requires a $1 \pm \varepsilon^2/3$ multiplicative approximation. For convenience, write $\varepsilon_0 = \varepsilon^2/3$. Again via Chebyshev's inequality,

$$\begin{aligned} \mathbb{P}[|X - ||p||_2^2| \geq \varepsilon||p||_2^2] &\leq \frac{Var(X)}{\varepsilon^2||p||_2^4} \\ &\leq \frac{1}{\varepsilon^2||p||_2^4}\left[\frac{2||p||_2^2}{s(s-1)} + \frac{4||p||_3^3}{s}\right] \\ &= \frac{2}{s(s-1)\varepsilon_0^2||p||_2^2} + \frac{4||p||_3^3}{\varepsilon_0^2 s||p||_2^4} \end{aligned}$$

Since $||p||_2^2 \geq 1/n$,

$$\frac{2}{s(s-1)\varepsilon_0^2||p||_2^2} \leq 1/6$$

for $s \geq 4\sqrt{n}/\varepsilon_0$. Since $||p||_3 \leq ||p||_2$, we have $||p||_3^3/||p||_2^4 \geq 1/||p||_2$. Since $||p||_2 \geq 1/\sqrt{n}$,

$$\frac{4||p||_3^3}{\varepsilon_0^2 s||p||_2^4} \leq \frac{4}{\varepsilon_0^2 s||p||_2} \leq 1/6$$

for $s \geq 24\sqrt{n}/\varepsilon_0^2$. Therefore, for $s = \lceil 24\sqrt{n}/\varepsilon_0^2 \rceil$ samples with probability at least 2/3, we obtain a $1 \pm \varepsilon_0$ factor approximation for $||p||_2^2$.

In the application to testing uniformity with respect to ℓ_1 distance we have $\varepsilon_0 = \varepsilon^2/3$ and hence $O(\varepsilon^{-4}\sqrt{n})$ samples suffice.

(Note that the upper bound $||p||_3^3/||p||_2^4 \leq 1/||p||_2 \leq \sqrt{n}$ that we used is asymptotically optimal. Consider a distribution which has probability $p_n = 1/\sqrt{n}$, $p_i = 1/n$ for $i \leq n - \sqrt{n}$ and $p_i = 0$ otherwise. This distribution has $||p||_2^2 \approx 2/n$ and $||p||_3^3 \approx 1/n^{3/2}$.)

Improvements and a Lower Bound The original ℓ_2 distance tester for uniformity is based on a tester due to Goldreich and Ron []. The version here and extension to ℓ_1 distance is due to Batu et al. []. Note that the 4-th power dependence on the inverse distance $1/\varepsilon$ is not optimal. The exact power is a bit less of an issue in property testing because unlike the streaming case, the comparison is with a polynomial in n rather than $\log n$. An asymptotically optimal dependence of $\Theta(\varepsilon^{-2}\sqrt{n})$ was shown by Paninski []. The basic idea similarly involves collisions but instead the algorithm extimates the distance based on the number of distinct samples. This avoids the large variance one can get if there are certain elements, such as in the example above where the probability of a triple collision is too large.

A lower bound of $\Omega(\sqrt{n})$ on the number of samples needed is not hard to show using a distribution related to the hard instances for element distinctness testing. Consider the distribution p that has probability $1/n$ for all elements larger than $2\varepsilon n$, but has probability $2/n$ for the first εn and the rest 0. This has $||p - U||_1 = 2\varepsilon$, but is $s = o(\sqrt{n})$ then the algorithm will not see any collisions with probability near 1.

Extensions One can extend this algorithm to one that tests the distance to any fixed known distribution q using the above algorithm for the uniform distribution: Group the elements into bins based on their probabilities so that every element has the same probability up to a $1 + \varepsilon$ factor, down to probabilities at most $\varepsilon/(n \log_2 n)$ say. (Elements with smaller probability occur too rarely in total probability to matter.) This gives $O(\log_n / \log(1 + \varepsilon)) = O(\varepsilon^{-1} \log n)$ bins. The algorithm will first apply the naive sampling algorithm to estimate the probability of each bin. Since there are only the sizes of $O(\varepsilon^{-1} \log n)$ bins, Chernoff bounds imply that all sizes of the corresponding bins for p can be estimated with small additive error using only $\varepsilon^{-1} \log^{O(1)} n$ samples. If any of these is too far from that of q, the algorithm will reject. Within each bin, the distribution is approximately uniform and the above test can be applied, provided that the bin has sufficiently large probability under q. Again, if the error on the bins is too large, the algorithm will reject.

If p and q are both given as input, then the sample complexity required is larger, $\varepsilon^{-\Theta(1)} n^{2/3}$ but part of the general idea is similar to that of the uniform case. Namely, one samples, elements from both distributions and compares the collisions within the p samples and within the q samples and compares this to the collisions between the p and q samples. The variance of this test is not good if one of the distributions contains some elements that occur too frequently. However, by first filtering out the high probability elements (those with probability $\Omega(n^{-2/3})$) and checking that those agree for the two distributions using the naive algorithm, Batu et al. derive the above bound.