ECS 189A Sublinear Algorithms for Big Data **Fall 2024**

Lecture 12: Identity Testing and Poissonization

Lecturer: Jasper Lee *Scribe: Avery Li*

1 Identity Testing (Generalization of Uniformity Testing)

We begin with an explicitly known distribution $\mathbf{q} = (q_1, \ldots, q_n)$ on $[n]$, given m i.i.d samples from $\mathbf{p}$ over $[n]$, we want to test if

- $\mathbf{p} = \mathbf{q}$
- $d_{TV}(\mathbf{p}, \mathbf{q}) \geq \epsilon$

with probability $\geq 2/3$. As it turns out, the sample complexity required is still $\Theta\left(\frac{\sqrt{n}}{\epsilon^2}\right)$.

2 Tester Construction

To solve this problem in traditional statistics, Pearson's χ^2 test is used:

$$\tilde{Z} = \sum_{i \in [n]} \frac{(N_i - mq_i)^2}{mq_i}$$

where N_i is the number of occurrences of domain element i, m is the sample complexity, and q_i is the probability mass of domain element under the known distribution $\mathbf{q}$. The issue with attempting to use and analyze $\tilde{Z}$ is that the terms can have a large variance.

Algorithm 12.1 Identity Tester

1. Draw $k \sim \text{Poi}(m)$ samples from $\mathbf{p}$
2. For each $i \in [n]$, let N_i be the number of times we see element i
3. Compute $A = \{i \in [n] \mid q_i \geq \frac{\epsilon}{50n}\}$
4. Compute $Z = \sum_{i \in A} \frac{(N_i - mq_i)^2 - N_i}{mq_i}$
5. Accept if $Z \leq \frac{m\epsilon^2}{10}$, otherwise reject

Intuitively, modifying the χ^2 statistic is fine because the difference is not that far from χ^2. We begin by examining the test statistic:

$$Z = \sum_{i \in A} \frac{(N_i - mq_i)^2 - N_i}{mq_i}$$

The algorithm accepts if $Z \leq m\epsilon^2/10$. First we examine the expectation of the Z when $\mathbf{p} = \mathbf{q}$,

$$\mathbb{E}[Z] = \sum_{i \in A} \frac{N_i}{mq_i} \Rightarrow \mathbb{E}[\tilde{Z}] = \sum_{i \in A} \frac{mp_i}{mq_i} = \sum_{i \in A} \frac{mq_i}{mq_i} = |A| \leq n$$

This modification of the χ^2 statistic is designed to better control variance as opposed to the traditional χ^2 test where variance cannot be controlled.
Example: Consider the following setting:

$$\mathbf{p} = \mathbf{q} : q_1 = 1 - \frac{1}{n}, q_i = \frac{1}{n(n-1)}$$

Take $m \ll n$ samples, with high probability we only observe elements $i \neq 1$ either 0 or 1 times. For these rare events:

$$\frac{(N_i - mq_i)^2}{mq_i} \approx \frac{N_i^2}{mq_i} = \begin{cases} 0 & \text{if } N_i = 0 \\ \Theta(n) & \text{if } N_i = 1 \text{ (even when } m \approx n) \end{cases}$$

While any individual element may not be sampled, for large enough m, one of these elements will be sampled, which implies high variance. Compare this to the modified statistic where for $N_i = 0, 1$ we get

$$\frac{(N_i - mq_i)^2 - N_i}{mq_i} \approx \frac{N_i^2 - N_i}{mq_i} = 0.$$

3 Poissonization (Poisson Sampling)

If a distribution is far from uniform, we should be able to detect the case using Z, through its mean difference from the uniform case, and bounding Z's variance to separate it from the uniform case. A key challenge is that $\{N_i\}$ are not independent because $\sum N_i = m$. This makes calculating $\text{Var}[Z]$ difficult, as we need to account for covariance and we cannot use tail bounds. Instead of drawing a fixed number of samples, we can instead use **Poissonization**:

1. Pick $k \sim \text{Poi}(m)$
2. Draw k samples from $\mathbf{p}$

We do not need to worry about the number of samples being too large because for large m, $\text{Poi}(m)$ is well-concentrated (Homework 1).

Proposition 12.2. *Suppose we draw* $\text{Poi}(m)$ *samples from* $\mathbf{p}$. *Then:*

1. $N_i \sim \text{Poi}(mp_i)$
2. $\{N_i\}$ *are independent*

This result is not immediately obvious and the proof will not be covered here. Also note that a Poissonised tester using Poisson samples can be simulated by a normal tester taking at most $2m$ samples, failing immediately when greater than $2m$ samples are made. This fails with at most $\text{poly}(1/m)$ more probability. This means we can run the standard tester without Poissonization.

4 Algorithm Analysis

Theorem 12.3. *Running Algorithm 12.1 on input Poi($m = O(\frac{\sqrt{n}}{\epsilon^2})$) samples, tests identity to* $\mathbf{q}$ *vs* ϵ*-far from* $\mathbf{q}$ *with probability* $\geq 2/3$.

By Proposition 12.2, N_i are independent Poi(mp_i). We have access to Z and want to test if $\mathbf{p}$ is ϵ-far away from $\mathbf{q}$. The general layout of the proof is to calculate and bound the expectation and variance of Z and establish a gap for the ϵ-far case for some constant probability.

Proposition 12.4.

$$\mathbb{E}[Z] = m \sum_{i \in A} \frac{(p_i - q_i)^2}{q_i} = m\chi^2(p_A || q_A)$$

$$\mathrm{Var}[Z] = \sum_{i \in A} \left[2\frac{p_i^2}{q_i} + 4mp_i(p_i - q_i)^2 \right]$$

Proof.

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{i \in A} \mathbb{E}\left[\frac{(N_i - mq_i)^2 - N_i}{mq_i}\right] \\ &= \sum_{i \in A} \frac{\mathbb{E}[N_i^2] - 2mq_i\mathbb{E}[N_i] + m^2q_i^2 - \mathbb{E}[N_i]}{mq_i}\end{aligned}$$

Observe that for Poisson random variables, $\mathbb{E} = \lambda$, $\mathrm{Var} = \lambda$, so we can further simplify with $\mathbb{E}[N_i] = mp_i$ and $\mathbb{E}[N_i^2] = mp_i + m^2p_i^2$.

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{i \in A} \frac{mp_i + m^2p_i^2 - 2m^2p_iq_i + m^2q_i^2 - mp_i}{mq_i} \\ &= m \sum_{i \in A} \frac{(p_i - q_i)^2}{q_i}\end{aligned}$$

For the proof of Var[Z], refer to Appendix A of arXiv:1507.05952. □

It now suffices to show there is a gap between the expectations between the $\mathbf{p} = \mathbf{q}$ case and ϵ-far case and that the variance is small enough to separate the two distributions with some constant probability based on the accept-reject criteria in Algorithm 12.1.

Lemma 12.5. *If* $\mathbf{p} = \mathbf{q}$, $\mathbb{E}[Z] = 0$. *If* $d_{TV}(\mathbf{p}, \mathbf{q}) \geq \epsilon$, $\mathbb{E}[Z] \geq \frac{1}{5}m\epsilon^2$

Proof. For $\mathbf{p} = \mathbf{q}$, note that the summand is 0. An additional claim needed is if $d_{TV}(\mathbf{p}, \mathbf{q}) \geq \epsilon$, then $d_{TV}(\mathbf{p}_A, \mathbf{q}_A) \geq \frac{1}{\sqrt{20}}\epsilon$. When $d_{TV}(\mathbf{p}, \mathbf{q}) \geq \epsilon$:

$$\begin{aligned}\chi^2(\mathbf{p}_A || \mathbf{q}_A) &\geq \left(\sum_{i \in A} \frac{(p_i - q_i)^2}{q_i}\right)\left(\sum_{i \in A} q_i\right) \\ &\geq \left(\sum_{i \in A} |p_i - q_i| \cdot \frac{\sqrt{q_i}}{\sqrt{q_i}}\right)^2 \\ &= 4d_{TV}^2(\mathbf{p}_A, \mathbf{q}_A) \\ &> \frac{1}{5}\epsilon^2.\end{aligned}$$

From this we get, $\mathbb{E}[Z] = m\chi^2(\mathbf{p}_A||\mathbf{q}_A) \geq \frac{m\epsilon^2}{5}$, and we have established an expectation gap. □

Lemma 12.6. *If there are enough samples, e.g.* $m \geq 10^{10}\frac{\sqrt{n}}{\epsilon^2}$ *samples:*

- $\mathbf{p} = \mathbf{q}$*:* $\mathrm{Var}[Z] \leq 4n \leq \frac{1}{400}m^2\epsilon^4$
- $\mathbf{p}$ *is* ϵ*-far from* $\mathbf{q}$*:* $\mathrm{Var}[Z] \leq \frac{1}{100}(\mathbb{E}[Z])^2$

The proof of the variance bound will not be covered here.

Proof. Proof of Theorem 12.3: Using Chebyshev's inequality, we can bound the probability of Z deviating from its expectation:

$$\mathbb{P}(Z > \mathbb{E}[Z] + \sqrt{3}\sqrt{\mathrm{Var}[Z]}) \leq \frac{1}{3}$$
$$\mathbb{P}(Z < \mathbb{E}[Z] - \sqrt{3}\sqrt{\mathrm{Var}[Z]}) \leq \frac{1}{3}$$

For $\mathbf{p} = \mathbf{q}$, by Lemma 12.5 and 12.6 we have that

$$\mathbb{E}[Z] + \sqrt{3}\sqrt{\mathrm{Var}[Z]} \leq \frac{1}{10}m\epsilon^2$$

Therefore, the probability that Algorithm 14.1 does not accept is less than $\frac{1}{3}$. When $d_{TV}(\mathbf{p}, \mathbf{q}) \geq \epsilon$, using Lemmas 12.5 and 12.6:

$$\mathbb{E}[Z] - \sqrt{3}\sqrt{\mathrm{Var}[Z]} \geq \left(1 - \frac{\sqrt{3}}{10}\right)\mathbb{E}[Z] \geq \frac{1}{10}m\epsilon^2$$

Therefore, the probability that Algorithm 14.1 does not reject in this case is less than $\frac{1}{3}$, so we are done. □

Techniques Used in Proof of Lemma 12.6

- Cauchy-Schwarz
- AM-GM inequality
- $||x||_1 \leq ||x||_2$ (relationship between L1 and L2 norms)