CSCI 1951-W Sublinear Algorithms for Big Data **Fall 2020**

Lecture 14: Identity Testing, Poissonisation, Closeness Testing

Lecturer: Jasper Lee *Scribe: Harman Suri*

1 Identity Testing Setting

Given an explicitly known distribution $\mathbf{q} = (q_1, ..., q_n)$ over $[n]$, and m i.i.d. samples from some distribution $\mathbf{p}$ (also over $[n]$), we want to test whether,

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

with probability at least 2/3.

Note that this strictly generalizes the uniformity testing problem, as uniformity testing is identity testing with $\mathbf{q}$ that is the uniform distribution over $[n]$.

We again are interested in the sample complexity m required. We will show, as in the uniformity testing case, that $\Theta(\frac{\sqrt{n}}{\epsilon^2})$ samples are required. Note that because uniformity testing requires $\Omega(\frac{\sqrt{n}}{\epsilon^2})$ samples, so does identity testing, so we only need to show that identity testing requires $O(\frac{\sqrt{n}}{\epsilon^2})$ samples.

2 Tester Construction

Motivation: In "traditional" statistics, you would attack this problem using Pearson's χ^2 test, which invovles computing the χ^2 statistic

$$\tilde{Z} = \sum_i \frac{(N_i - mq_i)^2}{mq_i}$$

where N_i is the number of occurrences of element $i \in [n]$, and $\mathbb{E}[N_i] = mq_i$.

The issue with using this test statistic is that it may have large variance, and therefore be difficult to analyze. So we will consider a modified χ^2 tester that will suppress the variance without changing the expectation in an uncontrolled manner.

Algorithm 14.1 *Identity Tester*

1. Take m samples from the unknown distribution $\mathbf{p}$, and compute N_i.
2. Compute $A = \{i : q_i \geq \frac{\epsilon}{50n}\}$
3. Compute $Z = \sum_{i \in A} \frac{(N_i - mq_i)^2 - N_i}{mq_i}$
4. Accept if and only if $Z \leq \frac{1}{10} m\epsilon^2$

Note: There are many other optimal identity tester constructions, and using a modified χ^2 test is only one of the possible approaches.

Example: Consider the distribution $\mathbf{q}$ defined by $q_1 = 1 - \frac{1}{n}$ and $q_i = \frac{1}{n(n-1)}$ for all $i \neq 1$, and consider the identity testing problem with the distribution $\mathbf{p} = \mathbf{q}$. With high probability, we will only observe each element i at most once if we take $O(n)$ samples (i.e. it is very unlikely to observe something that is not element 1). If we look at the original χ^2 statistic, and consider $m = O(n)$, we see that:

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

This contrasts with the new construction as follows (note that $N_i^2 - N_i = 0$ if $N_i = 1$):

$$\frac{(N_i - mq_i)^2 - N_i}{mq_i} \approx \frac{N_i^2 - N_i}{mq_i} = \begin{cases} 0 & N_i = 0 \\ 0 & N_i = 1 \end{cases}$$

So using the new construction in this case we have drastically reduced the variance of each individual summand of the computed statistic as compared to the χ^2 tester.

3 Poissonisation

Motivation: Note that the N_i in our statistic are dependent, as they must sum to m. This makes analysis difficult as we cannot use tail bounds and $\mathrm{Var}[Z]$ has annoying covariance terms. The solution to this dependence issue is to use **Poissonisation**.

Poissonisation: Instead of drawing m samples from distribution $\mathbf{p}$, we will draw $\mathrm{Poi}(m)$ many samples. That is, we first draw $k \leftarrow \mathrm{Poi}(m)$, and then draw k samples from $\mathbf{p}$.

Because the Poissonisation process produces a random number of samples k that we must draw from our distribution, we should be concerned if k could be too large (meaning would be trying to draw too many samples) with nontrivial probability. However, for large m, $\mathrm{Poi}(m)$ is well concentrated (see problem 5 on homework 1), so sampling a k that is too large (compared to m) should not be a concern. What we are really gaining from Poissonisation is captured in the following proposition:

Proposition 14.2 *Suppose we draw* $\mathrm{Poi}(m)$ *samples from* $\mathbf{p}$. *Then*

1. $N_i \leftarrow \mathrm{Poi}(mp_i)$
2. *All* N_i *are independent of each other.*

We will accept this proposition without proof, and with it we will be able to analyze $\mathrm{Var}[Z]$ without the covariance terms.

Note: A Poissonised tester taking $\mathrm{Poi}(m)$ samples can be simulated by a tester taking deterministically $2m$ samples, and failing with at most $\mathrm{poly}\,(\frac{1}{m})$ more probability. So we have a theoretically rigorous way to transform a Poissonised tester into a non-Poissonised one. In practice however, we can simply run the standard tester.

4 Upper Bound Analysis

Theorem 14.3 *Algorithm 14.1 on input* $\text{Poi}(m = O(\frac{\sqrt{n}}{\epsilon^2}))$ *samples, tests identity to* $\mathbf{q}$ *versus* ϵ*-far from* $\mathbf{q}$ *with probability at least 2/3*

Proof Idea: By Proposition 14.2, $N_i \overset{i.i.d.}{\leftarrow} \text{Poi}(mp_i)$. Therefore, we can calculate and bound $\mathbb{E}[Z]$ and $\text{Var}[Z]$, and then apply Chebyshev's inequality. We will first calculate $\mathbb{E}[Z]$ and $\text{Var}[Z]$ given in the proposition below. Note that $\chi^2(\mathbf{p}_A||\mathbf{q}_A)$ is the χ^2 divergence between $\mathbf{p}$ restricted to elements in A and $\mathbf{q}$ restricted to elements in A.

Proposition 14.4

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

$$\text{Var}[Z] = \sum_{i\in A} 2\frac{p_i^2}{q_i^2} + 4m\frac{p_i(p_i - q_i)^2}{q_i^2}$$

Proof. For $\mathbb{E}[Z]$*:*

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

We now use the fact that the $N_i \leftarrow \text{Poi}(mp_i)$, which means that $\mathbb{E}[N_i] = mp_i$. Furthermore, $\text{Var}[N_i] = mp_i$, so $\text{Var}[N_i] = mp_i = \mathbb{E}[N_i^2] - (\mathbb{E}[N_i])^2 = \mathbb{E}[N_i^2] - m^2p_i^2$. So $\mathbb{E}[N_i^2] = mp_i + m^2p_i^2$. This simplifies $\mathbb{E}[Z]$ from above as follows:

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

For $\text{Var}[Z]$*:* See Appendix A of arxiv:1507.05952. □

Given we now have exact expressions for the expectation and variance of Z, we want to show that the expectations in the identical case and the ϵ-far case have some gap, and that the variances in both cases are small enough so there is some constant probability concentration that is separated by the threshold required in Algorithm 14.1. We now bound the expectation and the variance in the two cases via the following Lemmas:

Lemma 14.5 *Bounding expectation.*

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

Lemma 14.6 *Bounding variance. If* $m \geq c\frac{\sqrt{n}}{\epsilon^2}$ *for sufficiently large constant* c*,*

- *If* $\mathbf{p} = \mathbf{q}$*, then* $\text{Var}[Z] \leq 4n \leq \frac{1}{400}m^2\epsilon^4$.

- *If* $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \geq \epsilon$, *then* $\mathrm{Var}[Z] \leq \frac{1}{100}(\mathbb{E}[Z])^2$.

We will first use these two lemmas to prove Theorem 14.3.

Proof of Theorem 14.3. Observe that by Chebyshev's inequality,

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

If $\mathbf{p} = \mathbf{q}$, then $\mathbb{E}[Z] + \sqrt{3}\sqrt{\mathrm{Var}[Z]} \leq \frac{1}{10}m\epsilon^2$ using Lemmas 14.5 and 14.6. Therefore, the probability that Algorithm 14.1 will not accept (refer back to the acceptance condition) for this $\mathbf{p}$ is less than or equal to $\frac{1}{3}$.

Similarly, if $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \geq \epsilon$, then $\mathbb{E}[Z] - \sqrt{3}\sqrt{\mathrm{Var}[Z]} \geq (1 - \frac{\sqrt{3}}{10})\,\mathbb{E}[Z] > \frac{1}{10}m\epsilon^2$, again using the above Lemmas. So the probability of failing to reject this $\mathbf{p}$ is also less than or equal to $\frac{1}{3}$. □

We now continue with the proofs of Lemmas 14.5 and 14.6.

Proof of Lemma 14.5. If $\mathbf{p} = \mathbf{q}$, the proof is trivial using Proposition 14.4 and noting that every summand will be 0. For the $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \geq \epsilon$ case, we use the result that $d_{\mathrm{TV}}(\mathbf{p}_A, \mathbf{q}_A) \geq \frac{1}{\sqrt{20}}\epsilon$, where $\mathbf{p}_A$ and $\mathbf{q}_A$ are again $\mathbf{p}$ and $\mathbf{q}$ restricted to elements of A. Intuitively, this is true because the set A constitutes only a tiny bit of probability mass in $\mathbf{q}$, and showing the statement is a good exercise. A proof can be found in Appendix B at arXiv:1507.05952.

We then recall that $\mathbb{E}[Z]$ was defined in terms of $\chi^2(\mathbf{p}_A||\mathbf{q}_A)$, which we can expand as follows:

$$\begin{aligned}
\chi^2(\mathbf{p}_A||\mathbf{q}_A) &\geq \left(\sum_{i\in A}\frac{(p_i - q_i)^2}{q_i}\right)\left(\sum_{i\in A} q_i\right) \\
&\geq \left(\sum_{i\in A}|p_i - q_i|\frac{\sqrt{q_i}}{\sqrt{q_i}}\right)^2 \quad \text{by Cauchy-Schwarz} \\
&= 4d_{\mathrm{TV}}^2(\mathbf{p}_A, \mathbf{q}_A) \\
&\geq \frac{1}{5}\epsilon^2 \quad \text{by the above result}
\end{aligned}$$

So $\mathbb{E}[Z] = m\chi^2(\mathbf{p}_A||\mathbf{q}_A) \geq \frac{1}{5}m\epsilon^2$, which shows Lemma 14.5. □

Proof of Lemma 14.6. We first make a claim to aid in the proof. The claim will be proved after it is applied.

Claim 14.7 $\mathrm{Var}[Z] \leq 4n + 9\sqrt{n}\,\mathbb{E}[Z] + \frac{2}{5}n^{1/4}(\mathbb{E}[Z])^{3/2}$

Using Claim 14.7, if $\mathbf{p} = \mathbf{q}$, then $\mathrm{Var}[Z] \leq 4n$. If $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \geq \epsilon$, then by the assumption that $m \geq c \cdot \frac{\sqrt{n}}{\epsilon^2}$) for a sufficiently large constant c, we have from Lemma 14.5 that $\mathbb{E}[Z] \geq \frac{1}{5}m\epsilon^2 \geq 4000\sqrt{n}$. Then, $\mathrm{Var}[Z] \leq O(\mathbb{E}[Z])^2 + O(\mathbb{E}[Z])^2 + O(\mathbb{E}[Z])^2 \leq \frac{1}{100}(\mathbb{E}[Z])^2$

Thus it remains to prove Claim 14.7.

Proof of Claim 14.7:

Recall from Proposition 14.4 that

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

We will bound each term separately. The first term can be expressed as:

$$2\sum_{i\in A}\frac{p_i^2}{q_i^2} = 2\sum_{i\in A}\frac{q_i^2}{q_i^2} + 2\sum_{i\in A}\left(\frac{(p_i-q_i)^2}{q_i^2} + 2\frac{(p_i-q_i)}{q_i}\right)$$

We then note that $2\frac{(p_i-q_i)}{q_i} = \sqrt{\frac{(p_i-q_i)^2}{q_i^2}\cdot 1}$, and so we can apply the AM-GM inequality (and use the fact that $\sum_i \frac{q_i^2}{q_i^2} \le n$) to obtain:

$$\begin{aligned}
2\sum_{i\in A}\frac{p_i^2}{q_i^2} &\le 2n + 2\sum_{i\in A}\left(\frac{(p_i-q_i)^2}{q_i^2} + \frac{(p_i-q_i)^2}{q_i^2} + 1\right)\\
&\le 4n + 4\sum_{i\in A}\frac{(p_i-q_i)^2}{q_i^2}\\
&\le 4n + \frac{200n}{\epsilon}\sum_{i\in A}\frac{(p_i-q_i)^2}{q_i} \text{ by the fact that } q_i \ge \frac{\epsilon}{50n}\\
&= 4n + \frac{200n}{\epsilon}\frac{\mathbb{E}[Z]}{m}\\
&\le 4n + \frac{1}{100}\sqrt{n}\,\mathbb{E}[Z] \text{ (for } m \text{ sufficiently large)}
\end{aligned}$$

Bounding the other term of $\mathrm{Var}[Z]$:

$$\begin{aligned}
4m\sum_{i\in A}\frac{p_i(p_i-q_i)^2}{q_i\cdot q_i} &\le 4m\sqrt{\sum_{i\in A}\frac{p_i^2}{q_i^2}}\sqrt{\sum_{i\in A}\frac{(p_i-q_i)^4}{q_i^2}} \text{ by Cauchy-Schwarz}\\
&\le 4m\sqrt{4n + \frac{1}{100}\sqrt{n}\,\mathbb{E}[Z]}\sqrt{\sum_{i\in A}\frac{(p_i-q_i)^4}{q_i^2}} \text{ by applying the bound from the first term}\\
&\le 4m(2\sqrt{n} + \frac{1}{10}n^{1/4}(\mathbb{E}[Z])^{1/2})(\sum_{i\in A}\frac{(p_i-q_i)^2}{q_i}) \text{ because } \|x\|_2 \le \|x\|_1\\
&= (8\sqrt{n} + \frac{2}{5}n^{1/4}(\mathbb{E}[Z])^{1/2})\,\mathbb{E}[Z]
\end{aligned}$$

Thus we have bounded both terms of $\mathrm{Var}[Z]$ within the bound given by Claim 14.7, which in turn proves Lemma 14.6. □

Note: In the identity testing setting, since we only need concentration in the constant probability sense, and we are able to bound the variance, Chebyshev's inequality alone is able to give us something tight. Also, applying some of our known concentration inequalities (e.g. Hoeffding's) would not work as each summand of our statistic Z is unbounded.

5 Closeness Testing

Given an two unknown distributions $\mathbf{p}, \mathbf{q}$ over $[n]$, and m *iid* samples from each of $\mathbf{p}$ and $\mathbf{q}$, we want to test whether,

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

with probability at least $2/3$.

Observation 14.8 *Identity testing reduces to closeness testing by simulating samples from known distribution* $\mathbf{q}$.

We state one optimal algorithm for the closeness testing problem.

Algorithm 14.9 (Closeness Tester)

1. *Get samples from* $\mathbf{p}, \mathbf{q}$ *and compute* X_i*, the number of times element* i *is seen from the* $\mathbf{p}$ *samples, and* Y_i*, the number of times element* i *is seen from the* $\mathbf{q}$ *samples*
2. *Compute* $Z = \sum_i \frac{(X_i - Y_i)^2 - X_i - Y_i}{X_i + Y_i}$
3. *Accept iff* $Z \leq \frac{1}{8} \frac{m^2}{m+n} \epsilon^2$

The following theorem claims that this algorithm is optimal for closeness testing.

Theorem 14.10 (Chan, Diakonikolas, Valiant, Valiant 2014) *Algorithm 14.9, on input* $m_1 = \mathrm{Poi}(m = O(\max(\frac{n^{2/3}}{\epsilon^{4/3}}, \frac{\sqrt{n}}{\epsilon^2})))$ *samples from* $\mathbf{p}$ *and* $m_2 = \mathrm{Poi}(m = O(\max(\frac{n^{2/3}}{\epsilon^{4/3}}, \frac{\sqrt{n}}{\epsilon^2})))$ *samples from* $\mathbf{q}$ *tests closeness of* $\mathbf{p}, \mathbf{q}$ *(versus* ϵ*-far) with probability at least* $2/3$*. Also, no algorithm can take* $o(\max(\frac{n^{2/3}}{\epsilon^{4/3}}, \frac{\sqrt{n}}{\epsilon^2})))$ *samples from both* $\mathbf{p}$ *and* $\mathbf{q}$ *and still succeed with probability at least* $2/3$.