COMS E6998-9: Algorithms for Massive Data (Spring'19) May 10, 2019

Identity Testing

Authors: *Weston Jackson (wjj2106), Jaewan Bahk (jb3621), Vu-Anh Phung (adp2161)*

Abstract

The field of distribution testing has evolved rapidly over the past two decades, gaining ever-increasing importance as data sets increase in size. In this paper, we survey recent breakthroughs for two important distribution testing subproblems: *identity testing* and *uniformity testing*. The identity testing problem asks us, given a fixed distribution p, how many samples from an unknown distribution q are needed to distinguish $p = q$ from $\|p - q\|_1 \geq \epsilon$? The uniformity testing problem is similar, asking us to solve this problem in the case where p is the uniform distribution. Our survey will cover a variety of sublinear time algorithms for solving these problems, including collision-based and coincidence-based testers, culminating in a tight instance-optimal tester from Valiant and Valiant.

Contents

1 Introduction

Analyzing and learning from massive amounts of data is of utmost importance in computer science, statistics, and many other related fields. While the mantra of bigger is better is often true when it comes to data sets, in many cases the amount of data can be so large that traditional algorithms for testing underlying distributions become infeasible. In these instances, we turn our attention to *distribution testing* algorithms. Distribution testing algorithms are sampling algorithms that allow us to learn specific properties of the data efficiently as opposed to learning an entire distribution. In many cases, recent advancements in distribution testing have led to significant reductions in the time complexity for testing properties of data sets and thus understanding their underlying distributions.

More formally, distribution testing algorithms ask us how many times do we need to draw elements from an unknown distribution q, over n elements, in order to test whether q has some property. One of the most important problems in distribution testing is the *identity testing* problem.

Definition 1 (Identity Testing Problem)**.** *Given a known distribution p, how many samples $x_1 \cdots x_m$ from an unknown distribution q are needed to distinguish $p = q$ from $\|p - q\|_1 \geq \epsilon$ with probability 2/3.*

The goal is to create an algorithm which solves the identity testing problem using as few samples as possible. The identity testing problem is a generalized form of the *uniformity testing* problem, which asks how many samples are needed when p is the uniform distribution.

One solution to the identity testing problem is via distribution learning. Distribution learning requires us to take i.i.d samples from the unknown distribution q until we can learn q to ϵ accuracy. From [1], it is known that we can learn q to sufficient accuracy with $O(\frac{n}{\epsilon^2})$ samples. Thus, one trivial solution to test whether p and q are identical would be to take $O(\frac{n}{\epsilon^2})$ samples to learn q, and then compute whether $\|p-q\|_1 \leq \epsilon$. The inefficiency with distribution learning is that learning q in order to determine if $\|p-q\|_1$ is not actually necessary. In particular, the solution doesn't exploit the distribution of p in any way. The key to property testing with less samples will be to compare properties of p and q which can be estimated with sublinear samples.

2 Sublinear Testing Background

2.1 Goldreich and Ron

Goldreich and Ron study property testing as it is applies to graphs [2]. Their input is a graph G on n vertices with bounded degree d. They then ask how many queries on G are needed to determine if G is an *expander* with probability 2/3. An expander graph is a graph in which the second eigenvalue of the adjacency matrix is at most $\lambda \in [0, 1]$. The goal is to reject if the graph is ϵ-far from having second eigenvalue at most λ' where $\lambda' = \lambda^{\alpha/O(1)}$ (we assume $\alpha < 0.5$ and $\lambda < \lambda'$). Goldreich and Ron's algorithm for expander graph testing is given in Algorithm 1.

For any starting vertex s, denote $p_{s,v}$ the probability that a random walk of length L starts at s and ends at v. If the algorithm accepts after $O(\frac{n^{0.5+\alpha}}{\epsilon^2})$ queries, Goldreich and Ron prove that with constant probability, the value of $\sum_{v\in[n]} p_{s,v}^2$ is within a $1 \pm \frac{n^{-\alpha/2}}{4}$ factor to a graph in which the collision probability is at most $\frac{1}{n}$. Thus, the ℓ_2 distance between the probability vector $(p_{s,v})_{v\in[n]}$ and the uniform probability vector is close. Note that in the case where $\alpha = 0$, the algorithm can determine if $\sum_{v\in[n]} p_{s,v}^2$ has ℓ_2 distance within a $(1 \pm \epsilon)$ factor of the uniform distribution with $O(\frac{\sqrt{n}}{\epsilon^2})$ samples.

Theorem 2. *Given an unknown distribution q over n, there is a test using $O(\frac{\sqrt{n}}{\epsilon^2}\log(1/\delta))$ samples that estimates $\|q\|_2$ to within a factor of $(1 \pm \epsilon)$ with probability $1 - \delta$*

Algorithm 1 Goldreich and Ron

```
function ExpanderGraphTester($G, \alpha, \lambda, \epsilon$)
    Set $L = \frac{1.5 \ln n}{\ln(\frac{1}{\lambda})}$
    for $t = \Theta(\frac{1}{\epsilon})$ times do:
        Select uniformly a start vertex $s$
        Perform $m = \Theta(\frac{n^{0.5+\alpha}}{\epsilon})$ random walks of length $L$ starting from $s$
        $C :=$ # collisions on endpoints of these $m$ walks
        if $C > \frac{1+0.5n^{\frac{-\alpha}{2}}}{n}\binom{m}{2}$ then
            return REJECT
        end if
    end for
    return ACCEPT
end function
```

We can use the ℓ_2 sampler from Goldreich and Ron for the purpose of uniformity testing. The ℓ_2 distance of the probability vector $(p_{s,v})_{v \in [n]}$ from the uniform probability vector is $\sum_v p_{s,v}^2 - \frac{1}{n}$. Furthermore, the ℓ_1 distance is bounded by $\sqrt{n}$ times the ℓ_2 distance.

$$\|x - y\|_1 \leq \sqrt{n}\|x - y\|_2$$

Thus, testing uniformity to $1 + \epsilon$ in ℓ_1 actually requires us to test within $\frac{\epsilon}{\sqrt{n}}$ in ℓ_2. With this goal in mind, Goldreich and Ron's strategy requires $O(\frac{\sqrt{n}}{\epsilon^4})$ total samples to solve the uniformity testing problem.

2.2 Batu et al.

In [3], Batu et al. use bucketing as well as the collision-based algorithm from Goldreich and Ron to reduce the identity testing problem to the problem of testing several approximately uniform distributions. Given a distribution p over R, they define $\{R_0 \cdots R_k\}$ as a partition of R into $k = O(\frac{\log n}{\epsilon})$ buckets. Each bucket R_i is defined as follows:

$$R_i = \{j \in [n] : \frac{(1+\epsilon)^{i-1}}{2n} \leq p_j \leq \frac{(1+\epsilon)^i}{2n}\}$$

Lemma 3. *Let p be an explicit distribution over R. Let U be the uniform distribution. Let $\{R_0 \cdots R_k\}$ be the bucketed partition of p into $O(\frac{\log n}{\epsilon})$ buckets. Then the restriction of p to the buckets of R_i is approximately uniform:*

$$\|Pr_p(R_i) - Pr_U(R_i)\|_1 \leq \epsilon$$

$$\|Pr_p(R_i) - Pr_U(R_i)\|_2 \leq \frac{\epsilon}{\sqrt{|R_i|}}$$

Since the restriction of p to any R_i is approximately uniform, we can use the ℓ_2 sampler from Goldreich and Ron to ensure the sample distribution is close to uniform in the $O(\frac{\log n}{\epsilon})$ buckets. In particular, we sample from q and test if q_{R_i} is close to uniform on each R_i. The full tester is given in Algorithm 2.

The algorithm requires the number of samples on each bucket be sufficient to determine $\|q_{R_i}\|^2$ to high accuracy via Goldreich and Ron. We then ensure that $\|q_{R_i}\|^2$ is close to uniform by rejecting if $\|q_{R_i}\|^2 > \frac{1+\epsilon^2}{|R_i|}$. Finally, the last $|p_R - q_R| > \epsilon$ check is needed to ensure q does not sample too many elements that are outside of $R_0 \cdots R_k$. Using a similar argument to Goldreich and Ron's uniformity

Algorithm 2 Batu et al.

```
function IdentityTester($p, q, \epsilon$)
    Partition (bucket) distribution $p$ into $k$ partitions, $R_1 \cdots R_k$
    Obtain $O(\sqrt{n}\epsilon^{-2}\log n)$ samples from $q$
    for partition $R_i$ s.t. $p_{R_i} \geq \frac{\epsilon}{k}$ do:
        if # collisions on $R_i < O(\frac{\sqrt{n}}{\epsilon^4})$ then
            return REJECT
        end if
        Estimate $\|q_{R_i}\|^2$ using Goldreich and Ron
        if $\|q_{R_i}\|^2 > \frac{1+\epsilon^2}{|R_i|}$ then
            return REJECT
        end if
    end for
    if $|p_R - q_R| > \epsilon$ then
        return REJECT
    end if
    return ACCEPT
end function
```

testing argument, Batu et al. show that this algorithm succeeds at differentiating $p = q$ from $\|p-q\|_1 \geq \epsilon$ with constant probability[1].

The sample complexity for the entire algorithm is $O(\sqrt{n}\log(n)poly(\epsilon^{-1}))$, which arises from the requirement to use Goldreich and Ron ℓ_2 estimator $O(\frac{\log n}{\epsilon})$ times.

Theorem 4. *There is an identity testing algorithm for any fixed p and unknown distribution q that requires $O(\sqrt{n}\log(n)poly(\epsilon^{-1}))$ samples.*

3 Paninski's Optimal Uniformity Tester

3.1 Upper Bound

In [4], Paninski proposes a tight $\Theta(\frac{\sqrt{n}}{\epsilon^2})$ uniformity tester that improves on Goldreich and Ron's $O(\frac{\sqrt{n}}{\epsilon^4})$ algorithm. Rather than look at the number of collisions, Paninski looks at the number of *coincidences* – the number of elements sampled exactly once (denoted by K_1). The basic idea is that deviations from uniformity will lead to more collisions, and hence less coincidences.

Let $\mathbb{E}_u[K_1]$ be the expected number of elements sampled exactly once for the uniform distribution. Let m be the number of samples. The test simply rejects if $\mathbb{E}_u[K_1] - \mathbb{E}_q[K_1] > T_\alpha$ for a threshold $T_\alpha = \frac{m^2\epsilon^2}{2n}$. The proof of correctness relies on bounding $\mathbb{E}_q[K_1]$ using the following lemma.

Lemma 5. *In the case where $\|U - q\|_1 \geq \epsilon$:*

$$\mathbb{E}_u[K_1] - \mathbb{E}_q[K_1] \geq \frac{m^2\epsilon^2}{n}(1 + O(m/n))$$

[1] The tester from Batu et al. can actually differentiate $\|p-q\|_1 \leq \Theta(\frac{\epsilon^3}{\sqrt{n}\log(n)})$ from $\|p-q\|_1 \geq \epsilon$, which is a stronger bound than is necessary for identity testing.

Algorithm 3 Paninski

```
function UniformityTester($(p, q, \epsilon)$)
    $m := O(\sqrt{n}\epsilon^{-2})$
    $T_\alpha := \frac{m^2\epsilon^2}{2n}$
    Obtain $m$ samples from $q$
    if $m(\frac{n-1}{n})^{m-1} - K_1 > T_\alpha$ then
        return REJECT
    end if
    return ACCEPT
end function
```

Consider an arbitrary distribution q. The probability that we sample a given element q_i exactly once over m samples is $\binom{m}{1}q_i(1-q_i)^{m-1}$. Thus, the expected number of elements sampled exactly once is

$$\mathbb{E}_q[K_1] = \sum_{i=1}^{n} \binom{m}{1} q_i(1-q_i)^{m-1}$$

For the uniform distribution, this works out to $\mathbb{E}_u[K_1] = m\left(\frac{n-1}{n}\right)^{m-1}$. Define $f : [0,1] \to \mathbb{R}$ as follows:

$$f(x) = x\left(1 - \left(\frac{n}{n-1}(1-x)\right)^{m-1}\right)$$

We can then rewrite

$$\mathbb{E}_u[K_1] - \mathbb{E}_q[K_1] = m\left(\frac{n-1}{n}\right)^{m-1} - \sum_{i=1}^{n}\binom{m}{1}q_i(1-q_i)^{m-1} = m\left(\frac{n-1}{n}\right)^{m-1}\sum_{i=1}^{n} f(q_i)$$

We want to bound $\sum_{i=1}^{n} f(q_i)$ from below, which can easily be done by applying Jensen's inequality if f is convex. Unfortunately this is not the case, so Paninski develops a new lower bound for f using a clever choice of a convex, symmetric function g that is strictly increasing on input $g(|x|)$:

$$f(x) \geq g\left(|x - \frac{1}{n}|\right) + f'\left(\frac{1}{n}\right)\left(x - \frac{1}{n}\right)$$

We can now apply Jensen's inequality:

$$\mathbb{E}_u[K_1] - \mathbb{E}_q[K_1] \geq m\left(\frac{n-1}{n}\right)^{m-1}\sum_{i=1}^{n} g\left(|q_i - \frac{1}{n}|\right) \geq m\left(\frac{n-1}{n}\right)^{m-1} ng(\epsilon/n)$$

Then after massaging the function g, the lower bound follows.

$$\mathbb{E}_u[K_1] - \mathbb{E}_q[K_1] \geq \frac{m^2\epsilon^2}{n}(1 + O(m/n))$$

Lemma 6.

$$Var_q(K_1) \leq \mathbb{E}_u[K_1] - \mathbb{E}_q[K_1] + O(m^2/n)$$

The key insight for the proof is to use the strong Efron-Stein inequality for bounding variance:

$$Var(S) \leq \frac{1}{2}\mathbb{E}[\sum_{j=1}^{m}(S - S^{(i)})^2]$$

where S is a function of random variables and $S^{(i)} = S(x_1 \cdots x_i' \cdots x_m)$ is S computed with an i.i.d copy of x_i'. From here, we substitute $S = K_1$ with x_i being the independent samples from p. We omit the rest of the proof as it just consists of bounding $\frac{1}{2}\mathbb{E}[\sum_{j=1}^{m}(S - S^{(i)})^2]$.

Let $T := \mathbb{E}_u[K_1] - K_1$ and $T_\alpha := \frac{m^2\epsilon^2}{2n}$, There are two cases via Chebyshev:

- When $p = q$, we have that $\mathbb{E}[T] = 0$ and $Var = O(m^2/n)$. By Chebyshev $T \geq T_\alpha$ with probability $O(\frac{m^2}{nT_\alpha^2}) = O(\frac{n}{\epsilon^4 m^2}) = O(1)$ when $m = \Theta(\frac{\sqrt{n}}{\epsilon^2})$.
- When $\|p - q\| \geq \epsilon$, Paninski shows that the z-score is large $\frac{\mathbb{E}[T]}{\sqrt{Var[T]}} = O(\frac{m^2\epsilon^2/n}{\sqrt{m^2/n}}) = O(1)$ when $m = \Theta(\frac{\sqrt{n}}{\epsilon^2})$.

Theorem 7. *There is a uniformity testing algorithm for any unknown distribution q that requires only $O(\frac{\sqrt{n}}{\epsilon^2})$ samples.*

3.2 Lower Bound

While it is easy to show that $\Omega(\sqrt{n})$ samples are needed for unifomirty testing (see Appendix B), a tight bound on ϵ was not proven until Paninski's paper. The proof of the tight bound first assumes the number of elements n is even. We then samples $n/2$ Bernoulli random variables $z_i \in \{-1, 1\}$. Consider the hard distribution q where:

$$q_i = \begin{cases} \frac{1+\epsilon z_{i/2}}{n}, i \text{ is even} \\ \frac{1-\epsilon z_{(i+1)/2}}{n}, i \text{ is odd} \end{cases}$$

Essentially, each pair of consecutive domain elements $2i - 1$ and $2i$ are assigned probabilities slightly deviating from the uniform probability, namely $\frac{1+\epsilon}{n}$ and $\frac{1-\epsilon}{n}$ or vice versa. Given this definition, it is easy to show that q is exactly ϵ-far from the uniform distribution. However, if we draw $m = o(\frac{\sqrt{n}}{\epsilon^2})$ samples from the uniform distribution and q, Paninski shows the statistical distance between the resulting two m-fold product distributions is small.

Let Q and U be the product distribution after drawing m samples from q and the uniform distribution respectively. The proof uses a method from Pollard [5]. Define

$$\Delta = \frac{dQ}{dU} = 2^{-n/2} \sum_{z \in \{-1,1\}^{n/2}} \prod_{j=1}^{m}(1 + G(x_j, z))$$

as the density of Q w.r.t. U after m samples where $G(x_j, z) = \epsilon z_{j/2}$ or $\epsilon z_{(j+1)/2}$ if the jth sample is even or odd. Paninski then substitutes ℓ_1 with ℓ_2 and bounds

$$\|U - Q\|_2 = (\mathbb{E}[(\Delta - 1)^2])^{1/2} \leq (e^{\frac{m^2\epsilon^4}{n}} - 1)^{1/2}$$

by expanding the inner $(\Delta - 1)^2$ term and exploiting the fact that $\mathbb{E}_u[G(x_j, z)] = 0$ (all the terms cancel which are not of the form $\mathbb{E}_u[G(x_j, z)G(x_j, z')]$). Thus, if $m = o(\frac{\sqrt{n}}{\epsilon^2})$, the expression on the RHS is further bounded away from a constant, which means that the ℓ_1 distance between two m-fold product distributions is arbitrarily small.

Theorem 8. *Any general uniformity testing algorithm requires $\Omega(\frac{\sqrt{n}}{\epsilon^2})$ samples.*

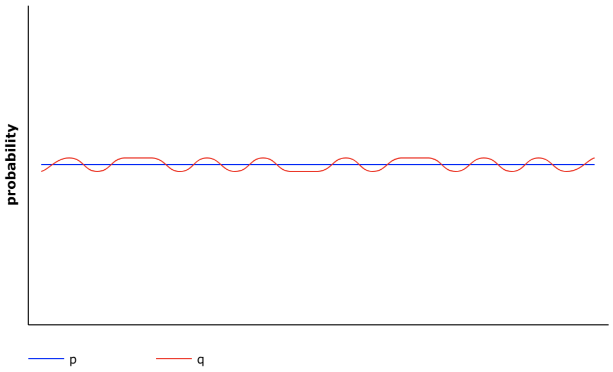


Figure 1: Example known distribution p and hard distribution q for Paninski's lower bound.

4 Valiant and Valiant's Instance Optimal Identity Tester

4.1 Intuition

While Paninski's lower bound established a tight bound on uniformity testing, lower bounds for identity testing remained an open question. However in 2014, Valiant and Valiant established a tight bound for identity testing in the *instance optimal* case [6].

Definition 9. *An instance-optimal identity testing algorithm is an algorithm that uses $f(p,\epsilon)$ samples to distinguish $p = q$ from $\|p - q\| \geq \epsilon$.*

Valiant and Valiant give a tight instance optimal identity testing algorithm that requires sample complexity upper bounded by $O(\frac{\|p\|_{2/3}}{\epsilon^2})$. Interestingly, the $2/3$ norm of the probability distribution p turns out to be a natural bound for identity testing. From the Paninski lower bound, we can see such a bound is tight when p is the uniform distribution:

$$O(\frac{\|U\|_{2/3}}{\epsilon^2}) = O(\frac{(\sum_{i=1}^{n} \frac{1}{n}^{2/3})^{3/2}}{\epsilon^2}) = O(\frac{(\frac{n}{n^{2/3}})^{3/2}}{\epsilon^2}) = O(\frac{\sqrt{n}}{\epsilon^2})$$

In the case where p is not the uniform distribution, the result immediately yields improved asymptotic complexity from the $O(\frac{\sqrt{n}\log n}{\epsilon^{O(1)}})$ algorithm from Batu et al.:

$$O(\frac{\|p\|_{2/3}}{\epsilon^2}) \leq O(\frac{\|U\|_{2/3}}{\epsilon^2}) = O(\frac{\sqrt{n}}{\epsilon^2})$$

Furthermore, in the case where p has large probability concentrated on just a few elements, the improvement is even greater.

4.2 Holder Inequality

Cauchy-Schwarz and Holder inequalities play an important role in the analysis of Valiant and Valiant's identity testing algorithm. They note that inequalities of the form

$$\prod_i (\sum x_j^{a_i} y_j^{b_i})^{c_i} \geq 1$$

for $(a)_i$, $(b)_i$, $(c)_i$ are often proven via trial and error. Valiant and Valiant show that such inequalities hold only when inequalities are expressible as a product of the following two forms for $\lambda \in [0,1]$:

- Holder inequalities

$$(\sum_j x_j^{a'} y_j^{b'})^{\lambda}(\sum_j x_j^{a''} y_j^{b''})^{1-\lambda} \geq \sum_j x_j^{\lambda a'+(1-\lambda)a''} y_j^{\lambda b'+(1-\lambda)b''}$$

- L_p monotonicity inequalities

$$(\sum_j x_j^a y_j^b)^{\lambda} \leq \sum_j x_j^{\lambda a} y_j^{\lambda b}$$

We omit the proof for conciseness.

4.3 χ^2-test

Valiant and Valiant's tester relies on an altered version of the classical χ^2-test. In 1900, Pearson's χ^2-test was the first statistical hypothesis test to measure the goodness of fit of a sample distribution to some true distribution p [7]. Assuming we draw k i.i.d samples X from q, and want to determine whether these samples match some distribution p, the χ^2-test asks us to calculate the test statistic

$$\sum_i \frac{(X_i - kp_i)^2}{p_i}$$

which compares the expected and observed frequencies of each element p_i in the known distribution. If the test statistic is small enough, meaning the difference between the expected and observed events do not differ by too much, we output $p = q$.

Unfortunately, in the case of rare events, the χ^2-test has large variance. Consider the distribution where $Pr(p_1) = 1 - \frac{1}{n}$ and the remaining $\frac{1}{n}$ probability is split across n elements with probability $\frac{1}{n^2}$. The remaining elements are extremely unlikely, yet on expectation one will appear every n samples. The appearance of a $\frac{1}{n^2}$ event assuming n samples will contribute

$$\frac{(1 - n\frac{1}{n^2})^2}{\frac{1}{n^2}} = \Omega(n^2)$$

to the calculation of the statistic. Thus, Pearson's statistic places a substantial amount of weight on rare events, and requires many samples in order to distinguish the sample and known distribution.

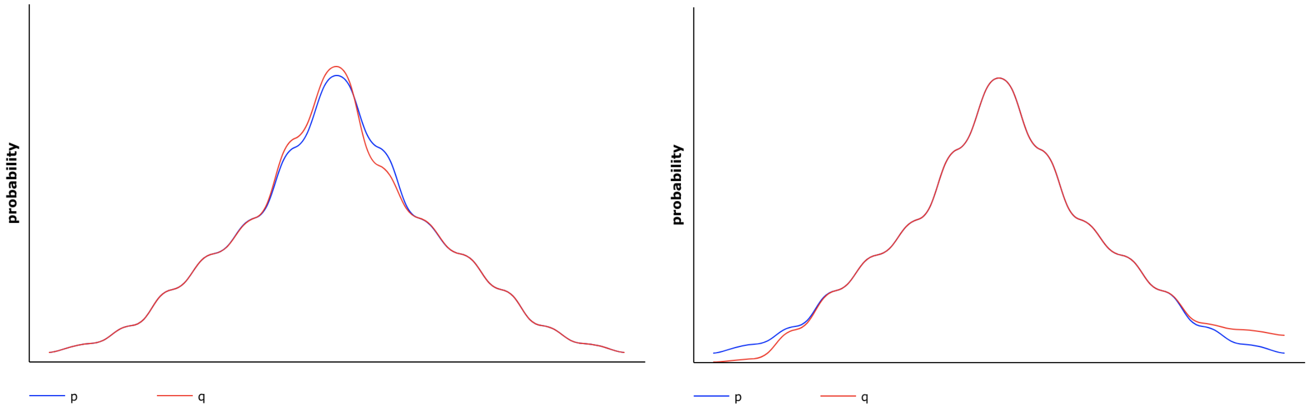
Valiant and Valiant alter Pearson's statistic in two ways to get an improved tester. They replace $1/p_i$ with $1/p_i^{2/3}$ in the denominator, then subtract $X_i/p_i^{2/3}$ from each term in the summation:

$$\sum_i \frac{(X_i - kp_i)^2 - X_i}{p_i^{2/3}}$$

Replacing $1/p_i$ with $1/p_i^{2/3}$ is important as it reduces the weight of rare events. Similarly, subtracting $X_i/p_i^{2/3}$ from each term means that rare events that appear zero or one times contribute less to the statistic. Consider the previous case where a $p_i = \frac{1}{n^2}$ event appears zero or one times in n samples:

$$\frac{(X_i - kp_i)^2 - X_i}{p_i^{2/3}} \approx (X_i^2 - X_i)p_i^{-2/3} = 0$$

Since the probability a $\frac{1}{n^2}$ event appears twice in $O(n)$ samples is extremely rare, Valiant and Valiant's tester has far less variance in these cases.



(a) Case 1: ϵ-discrepancy in common elements.

(b) Case 2: ϵ-discrepancy in uncommon elements.

Figure 2: The two cases for Valiant and Valiant's instance-optimal identity tester.

4.4 Upper Bound

In proving the sample complexity upper bound for their tester, Valiant and Valiant show that their tester only needs $O(\max\{\frac{1}{\epsilon}, \frac{\|p^{-m}_{-\epsilon/16}\|_{2/3}}{\epsilon^2}\})$ samples to achieve constant probability of success. The notation $p^{-m}_{-\epsilon}$ implies that if the elements of p are in sorted order based on probability mass, then we remove the smallest elements that have probability mass summing to ϵ and the element with maximum probability mass, p_m. Since $\frac{1}{\epsilon} < \frac{\|p\|_{2/3}}{\epsilon^2}$, this implies that

$$O(\max\{\frac{1}{\epsilon}, \frac{\|p^{-m}_{-\epsilon/16}\|_{2/3}}{\epsilon^2}\}) \leq O(\frac{\|p\|_{2/3}}{\epsilon^2})$$

Rather than consider collisions or coincidences, Valiant and Valiant's tester considers two cases. The first case is for detecting when the ϵ-discrepancy between p and q is concentrated in the common elements of p. Since the common elements occur with reasonable probability, Valiant and Valiant show that their χ^2-tester will have low variance. The second case is for detecting when the ϵ-discrepancy is concentrated in the rare elements of p, in particular the least likely elements with $O(\epsilon)$ total probability mass. In this case, although there can many rare elements, the effect of the discrepancy will be more obvious and will be detectable with $O(1/\epsilon)$ samples.

Assume that the domain elements of p are sorted in increasing order of probability, and let s be the largest integer such that $\sum_{i<s} p_i < \frac{\epsilon}{8}$. We also denote m as the index with maximum probability in p. The tester is given in Algorithm 4.

For analysis, we denote the statistic calculated on each common element i as χ_i:

$$\chi_i = \frac{(X_i - kp_i)^2 - X_i}{p_i^{2/3}}$$

It is straightforward to then calculate the expectation and variance of the terms for each χ_i:

$$\mathbb{E}[\chi_i] = \frac{k^2(p_i - q_i)^2}{p_i^{2/3}}$$

$$Var[\chi_i] = \frac{2k^2q_i^2 + 4k^3q_i(p_i - q_i)^2}{p_i^{4/3}}$$

Algorithm 4 Valiant and Valiant

```
function IdentityTester($(p, q, \epsilon)$)
    Obtain $Poi(c\max\{\frac{1}{\epsilon}, \frac{\|p^{-m}_{-\epsilon/16}\|_{2/3}}{\epsilon^2}\})$ samples from $q$
    if $\sum_{i\geq s, i\neq m}[(X_i - kp_i)^2 - X_i]p_i^{-2/3} > 4k\|p^{-m}_{\geq s}\|^{1/3}_{2/3}$ then
        return REJECT
    end if
    if $\sum_{i<s} X_i > \frac{3}{16}\epsilon k$ then
        return REJECT
    end if
    return ACCEPT
end function
```

The proof of correctness for the upper bound relies on the following key lemma.

Lemma 10 (Key Lemma)**.** *Let s be the largest integer such that $\sum_{i<s} p_i < \frac{\epsilon}{8}$. For sufficient $k = \Omega(\frac{\|p^{-m}_{-\epsilon/16}\|_{2/3}}{\epsilon^2})$ samples when $\epsilon_1 = \Omega(\epsilon)$ of the discrepancy falls above s (common elements), then for any $c_1 \geq 1$:*

$$c_1 Var[\sum_{i\geq s, i\neq m} \chi_i] < \mathbb{E}[\sum_{i\geq s, i\neq m} \chi_i]^2$$

or equivalently

$$c_1 \sum_{i\geq s, i\neq m} \frac{Var[\chi_i]}{k^4} < (\sum_{i\geq s, i\neq m} \Delta_i^2 p_i^{-2/3})^2$$

where $\Delta_i = |p_i - q_i|$

Proof. It can be shown that the number of samples $k \geq \frac{1}{c}\max\{\frac{\|p^{-m}_{\geq s}\|^{1/3}_{2/3}}{p_s^{1/3}\epsilon_1}, \frac{\|p^{-m}_{\geq s}\|_{2/3}}{\epsilon_1^2}\}$. By triangle inequality, $\frac{q_i}{k} \leq c(p_i\frac{\epsilon_1^2}{\|p^{-m}_{\geq s}\|_{2/3}} + \Delta_i\frac{p_s^{1/3}\epsilon_1}{\|p^{-m}_{\geq s}\|^{1/3}_{2/3}})$. Expanding $\sum_{i\geq s, i\neq m}\frac{Var[\chi_i]}{k^4}$:

$$c_1 \sum_{i\geq s, i\neq m} \frac{Var[\chi_i]}{k^4} \leq \sum_{i\geq s, i\neq m} [\underbrace{p_i^{2/3}\frac{\epsilon_1^4}{\|p^{-m}_{\geq s}\|^2_{2/3}}}_{(1)} + \underbrace{\Delta_i p_i^{-1/3}\frac{p_s^{1/3}\epsilon_1^3}{\|p^{-m}_{\geq s}\|^{4/3}_{2/3}}}_{(2)} + \underbrace{\Delta_i^2 p_i^{-4/3}\frac{p_s^{2/3}\epsilon_1^2}{\|p^{-m}_{\geq s}\|^{2/3}_{2/3}}}_{(3)}$$

$$+ \underbrace{\Delta_i^2 p_i^{-1/3}\frac{\epsilon_1^2}{\|p^{-m}_{\geq s}\|_{2/3}}}_{(4)} + \underbrace{\Delta_i^3 p_i^{-4/3}\frac{p_s^{1/3}\epsilon_1}{\|p^{-m}_{\geq s}\|^{1/3}_{2/3}}}_{(5)}]$$

We can bound each term by $(\sum_{i\geq s, i\neq m}\Delta_i^2 p_i^{-2/3})^2$ separately. The idea is to bound each term using $\epsilon_1 \leq \|\Delta^{-m}_{\geq s}\|_1$ and then some form of Cauchy-Schwarz inequality:

$$\frac{\epsilon_1^2}{\|p^{-m}_{\geq s}\|^{2/3}_{2/3}} \leq \frac{\|\Delta^{-m}_{\geq s}\|_1^2}{\|p^{-m}_{\geq s}\|^{2/3}_{2/3}} \leq \sum_{i\geq s, i\neq m}\Delta_i^2 p_i^{-2/3}$$

- (1) $\sum_{i\geq s, i\neq m} p_i^{2/3} = \|p^{-m}_{\geq s}\|^{2/3}_{2/3}$, then Cauchy-Schwarz (squared)

- (2) $\sum_{i\geq s,i\neq m} \Delta_i p_i^{-1/3} \leq \frac{\epsilon_1}{p_s^{1/3}}$, then Cauchy-Schwarz (squared)
- (3) $p_i^{-4/3} p_s^{2/3} \leq \|p_{\geq s}^{-m}\|_{2/3}^{-2/3}$, $\epsilon_1^2 \leq \|\Delta_{\geq s}^{-m}\|^2$, then Cauchy-Schwarz (squared)
- (4) Norm inequality, then Holder Inequality, to bound

  $$\sum_{i\geq s,i\neq m} \Delta_i^2 p_i^{-1/3} \leq (\sum_{i\geq s,i\neq m} \Delta_i^{4/3} p_i^{-2/9})^{3/2} \leq (\sum_{i\geq s,i\neq m} \Delta_i^2 p_i^{-2/3})\|p_{\geq s}^{-m}\|_{2/3}^{1/3}$$

  multiply LHS and RHS by Cauchy-Schwarz inequality, divide both sides by $\|p_{\geq s}^{-m}\|_{2/3}^{1/3}$.
- (5) Norm inequality, then Holder Inequality, to show

  $$\sum_{i\geq s,i\neq m} \Delta_i^3 p_i^{-4/3} \leq (\sum_{i\geq s,i\neq m} \Delta_i^2 p_i^{-8/9})^{3/2} \leq (\sum_{i\geq s,i\neq m} \Delta_i^2 p_i^{-2/3})^{3/2} p_s^{-1/3}$$

  multiply boths sides by square root of Cauchy-Schwarz inequality.

□

Theorem 11. *Valiant and Valiant's identity testing algorithm distinguishes $p = q$ from $\|p-q\|_1 \geq \epsilon$ with probability $2/3$ using $O(\max\{\frac{1}{\epsilon}, \frac{\|p_{-\epsilon/16}^{-m}\|_{2/3}}{\epsilon^2}\})$ samples.*

Proof. Recall that s is largest integer such that $\sum_{i<s} p_i < \frac{\epsilon}{8}$. The tester rejects in two cases:

1. $\sum_{i\geq s,i\neq m} \chi_i > 4k\|p_{\geq s}^{-m}\|_{2/3}^{1/3}$
2. $\sum_{i<s} X_i > \frac{3}{16}\epsilon k$

To prove that this tester works with constant probability, we need to show that we pass both checks in the case where $p = q$, and fail at least one check in the case where $\|p-q\|_1 \geq \epsilon$. Note when $\|p-q\|_1 \geq \epsilon$, at most $\frac{\epsilon}{2}$ discrepancy can be in p_m. Thus, $\Omega(\epsilon)$ fraction of the discrepancy either falls between p_s and p_m or below p_s. We show when a significant portion of this discrepancy is between p_s and p_m, case one rejects, otherwise case two rejects.

- $p = q \implies$ Case 1 passes:
  The expectation of the tester is $\sum_i \mathbb{E}[\chi_i] = \frac{k^2(p_i-q_i)^2}{p_i^{2/3}} = 0$ when $p = q$. The variance is $\sum_i Var[\chi_i] = \sum_i \frac{2k^2q_i^2+4k^3q_i(p_i-q_i)^2}{p_i^{4/3}} = 2k^2\|p_{\geq s}^{-m}\|_{2/3}^{2/3}$. By Chebyshev, the tester is less than $4k\|p_{\geq s}^{-m}\|_{2/3}^{1/3}$ with probability 7/8.
- $p = q \implies$ Case 2 passes:
  The probability we draw $\sum_{i<s} X_i$ elements $< s$ is distributed as $Poi(\frac{k\epsilon}{8})$ with expectation and variance $\frac{\epsilon k}{8}$. The probability this exceeds $\frac{3\epsilon k}{16}$ is less than 1/8 by Chebyshev for sufficient k.
- $\|p-q\|_1 \geq \epsilon \implies$ Case 2 fails when $\|(p-q)_{<s}^{-m}\| \geq \frac{3\epsilon}{8}$
  By assumption $\|p_{<s}\| < \frac{\epsilon}{8}$ and thus $\|q_{<s}\| \geq \frac{\epsilon}{4}$. Then the probability we draw $\sum_{i<s} X_i$ elements $< s$ is distributed as $Poi(\frac{k\epsilon}{4})$ with expectation and variance $\frac{\epsilon k}{4}$. The probability this exceeds $\frac{3\epsilon k}{16}$ is at least 7/8 by Chebyshev for sufficient k.

- $\|p - q\|_1 \geq \epsilon \implies$ Case 1 fails when $\|(p-q)^{-m}_{\geq s}\| \geq \frac{\epsilon}{8}$:
  Apply the key lemma. Thus, we have that $cVar[\chi_i] \leq \mathbb{E}[\chi_i]^2$ for any c and sufficient k. The variance is minimized when $p = q$, therefore is at least $c2k^2\|p^{-m}_{\geq s}\|^{2/3}_{2/3}$. The key lemma implies expectation is at least $\sqrt{2c}k\|p^{-m}_{\geq s}\|^{1/3}_{2/3}$. By Chebyshev, for sufficient choice of k, the tester is more than $4k\|p^{-m}_{\geq s}\|^{1/3}_{2/3}$ with probability at least 7/8.

□

4.5 Lower Bound

Valiant and Valiant show that their tester is optimal by constructing a hard distribution from a distribution over distributions Q_ϵ. The strategy is similar to Paninski's lower bound: we construct a distribution q^* from Q_ϵ that is ϵ-far from p, yet k samples from a random distribution q^* will be close to k samples from p via ℓ_1. Additionally, like Paninski, Valiant and Valiant choose q^* to be a random perturbation of the known distribution. However, to show that k samples from p and k samples from q^* are close, Valiant and Valiant rely on the Hellinger distance rather than working with ℓ_2 distance.

Definition 12 (Hellinger Distance)**.** *The Hellinger distance $H(p,q)$ between two distributions p and q is the following:*

$$H(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$$

The Hellinger distance is crucial to the lower bound as it bounds the ℓ_1 distance while its square is subadditive on product distributions. The idea will be to bound the ℓ_1 distance by summing over the squared Hellinger distances per coordinate. The bound relies on the following lemma:

Lemma 13. $H(Poi(\lambda), Poi(\lambda \pm \epsilon)) \leq O(\frac{\epsilon^2}{\lambda})$

The lemma is proved via properties of the Hellinger distance and Poisson distribution. We omit the details for conciseness. Valiant and Valiant then use it to prove the main theorem.

Theorem 14. *Given a distribution p and $\epsilon_i \in [0, p_i]$, draw q^* from Q_ϵ as follows*

$$q^*_i = p_i \pm \epsilon_i$$

then normalize q^* to be a distribution. It takes $k \geq \Omega((\sum_i \frac{\epsilon_i^4}{p_i^2})^{-1/2})$ samples to distinguish p from q^* with success probability $2/3$, and q^* is $\min\{(\sum_i \epsilon_i) - \max_i \epsilon_i, \frac{1}{2}\sum_i \epsilon_i\}$-far from p.

Proof. Let p_k and q^*_k denote p and q^* after sampling k times. We bound the ℓ_1 distance between p and q^* after k samples via the Hellinger distance. We can then bound the Hellinger distance using its subadditive property and the previous lemma:

$$\|p_k - q^*_k\|_1 \leq H(p_k, q_k) \leq \sqrt{\sum_i H(Poi(kp_i), Poi(k[p_i \pm \epsilon_i]))^2} \leq kO(\sum_i \frac{\epsilon_i^4}{p_i^2})^{1/2}$$

Thus, if $k = o((\sum_i \frac{\epsilon_i^4}{p_i^2})^{-1/2})$, the distance between p_k and q^*_k is arbitrarily small.

We next show that p and q^* are $\min\{(\sum_i \epsilon_i) - \max_i \epsilon_i, \frac{1}{2}\sum_i \epsilon_i\}$-far. We use the fact that the mass of q^* that is removed via normalization is distributed as $\sum_i \pm\epsilon_i$. Thus, the ℓ_1 distance is at least as large as sampling from $\sum_i \epsilon_i - |\sum_i \pm\epsilon_i|$. It suffices to show that $|\sum_i \pm\epsilon_i| \leq \max\{\max_i \epsilon_i, \frac{1}{2}\sum_i \epsilon_i\}$ with probability 1/2.

Let ϵ_1 be the largest ϵ_i value. If $\epsilon_1 \geq \frac{1}{2}\sum_i \epsilon_i$, then the randomness of choosing $\pm\epsilon_1$ implies $|\sum_i \pm\epsilon_i| \leq \epsilon_1$ with probability 1/2. For the case when $\epsilon_1 < \frac{1}{2}\sum_i \epsilon_i$, we consider the first element ϵ_j in sorted order for which it would be possible $|\frac{1}{2}\sum_{i<j} \pm\epsilon_i| + \epsilon_j$ exceeds $\frac{1}{2}\sum_i \epsilon_i$. Valiant and Valiant show the remaining elements after e_j will yield a sum at most $\frac{1}{2}\sum_i \epsilon_i$ with probability 1/2. In either case, $|\sum_i \pm\epsilon_i| \leq \max\{\max_i \epsilon_i, \frac{1}{2}\sum_i \epsilon_i\}$, and thus with probability 1/2, q^* is $\min\{(\sum_i \epsilon_i) - \max_i \epsilon_i, \frac{1}{2}\sum_i \epsilon_i\}$-far from p. □

Corollary 15. *There exists c such that for any $\epsilon \in (0,1)$ and known p, no tester can distinguish $p = q^*$ from $\|p - q^*\|_1 \geq \epsilon$ with probability $\geq 2/3$ with less than $c\max\{\frac{1}{\epsilon}, \frac{\|p_{-\epsilon}^{-m}\|_{2/3}}{\epsilon^2}\}$ samples.*

Proof. Let α be the value for which $\frac{1}{2}\sum_{i\neq m} \min\{p_i, \alpha p_i^{2/3}\} = \epsilon$. Then, let $\epsilon_i = \min\{p_i, \alpha p_i^{2/3}\}$. We can verify q^* and p are ϵ-far apart:

$$\|p - q^*\|_1 = \min\{(\sum_i \epsilon_i) - \max_i \epsilon_i, \frac{1}{2}\sum_i \epsilon_i\} \geq \min\{2\epsilon - \epsilon, \epsilon\} = \epsilon$$

Next, sort the p_i in ascending order and consider the largest s where $\sum_{i<s} p_i \leq \epsilon$. For all $p_i \geq p_s$, it can be shown that that $\min\{p_i, \alpha p_i^{2/3}\} = \alpha p_i^{2/3}$. This allows us to establish the following inequality:

$$\alpha \sum_{i=s}^{m-1} p_i^{2/3} = \sum_{i=s}^{m-1} \min\{p_i, \alpha p_i^{2/3}\} \leq \sum_{i\neq m} \min\{p_i, \alpha p_i^{2/3}\} \leq 2\epsilon$$

which implies $\alpha \geq 2\epsilon \|p_{\geq s}^{-m}\|_{2/3}^{-2/3}$. Thus, with this bound on α, we can show the number of samples needed to distinguish q^* and p is at least $\Omega(\frac{\|p_{\geq s}^{-m}\|_{2/3}}{\epsilon^2})$. Using the previous lemma:

$$\begin{aligned}
k &\geq \Omega((\sum_i \frac{\epsilon_i^4}{p_i^2})^{-1/2}) \\
&= \Omega((\sum_{i\neq m} \frac{\min\{p_i, \alpha p_i^{2/3}\}^4}{p_i^2})^{-1/2}) && \epsilon_i = \min\{p_i, \alpha p_i^{2/3}\} \\
&\geq \Omega((\alpha^3 \sum_{i\neq m} \min\{p_i, \alpha p_i^{2/3}\})^{-1/2}) \\
&= \Omega((\alpha^3 \epsilon)^{-1/2}) && \frac{1}{2}\sum_{i\neq m} \min\{p_i, \alpha p_i^{2/3}\} = \epsilon \\
&\geq \Omega((\epsilon^4 \|p_{\geq s}^{-m}\|_{2/3}^{-2})^{-1/2}) && \alpha \geq 2\epsilon \|p_{\geq s}^{-m}\|_{2/3}^{-2/3} \\
&= \Omega(\frac{\|p_{\geq s}^{-m}\|_{2/3}}{\epsilon^2})
\end{aligned}$$

□

The result implies that Valiant and Valiant's tester requiring $\Theta(\frac{\|p_{\geq s}^{-m}\|_{2/3}}{\epsilon^2})$ samples is tight.

5 Appendix

5.1 A

	Tester	Strategy	Complexity	Notes
Goldreich and Ron	Uniformity	Collisions	$O(\frac{n}{\epsilon^4})$	
Batu et al.	Identity	Collisions	$O(\frac{n \log n}{\epsilon^{O(1)}})$	
Paninski	Uniformity	Coincidences	$\Theta(\frac{n}{\epsilon^2})$	$\epsilon = \Omega(1/n^{1/4})$
Valiant and Valiant	Identity	χ^2-test	$\Theta(\frac{\|p\|_{2/3}}{\epsilon^2})$	

5.2 B

We can show $\Omega(\sqrt{n})$ samples are required for uniformity testing. Consider the distribution:

$$q = \begin{cases} \frac{1}{n}, \forall i \geq 2\epsilon n \\ \frac{2}{n}, \forall i \leq \epsilon n \\ 0, \text{ otherwise} \end{cases}$$

It is easy to verify that q is at least ϵ far from uniform. With $o(\sqrt{n})$ samples and assuming $\epsilon < 0.5$, we don't expect to see any collisions when sampling from the uniform distribution and q:

$$E[\# \text{ collisions}] = \|U\|_2^2 \binom{o(\sqrt{n})}{2} = \frac{1}{n} o(n) = o(1)$$

$$E[\# \text{ collisions}] = \|q\|_2^2 \binom{o(\sqrt{n})}{2} < \frac{2}{n} o(n) = o(1)$$

In either case there are no collisions on expectation, and thus $\Omega(\sqrt{n})$ samples are required.

References

[1] Luc Devroye and Gabor Lugosi. *Combinatorial Methods in Density Estimation.* Springer Series in Statistics, New York, NY, USA, 2001.

[2] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. In *Technical Report TR00-020, Electronic Colloquium on Computational Complexity*, 2000.

[3] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, Oct 2001.

[4] L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theor.*, 54(10):4750–4755, October 2008.

[5] David Pollard. Asymptopia. http://www.stat.yale.edu/ pollard/Books/Asymptopia, 2003. Manuscript.

[6] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, pages 51–60, Washington, DC, USA, 2014. IEEE Computer Society.

[7] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series*, 50(302):157–175, 1900.