# Sublinear algorithms for outlier detection and generalized closeness testing

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh

ECE UCSD, {jacharya, ashkan, alon, asuresh}@ucsd.edu

***Abstract*—*Outlier detection* is the problem of finding a few different distributions in a set of mostly identical ones. *Closeness testing* is the problem of deciding whether two distributions are identical or different. We relate the two problems, construct a sub-linear generalized closeness test for unequal sample lengths, and use this result to derive a sub-linear *universal outlier detector*. We also lower bound the sample complexity of both problems.**

## I. Introduction

Let $p_1, p_2, \ldots, p_m$ be $m$ unknown distributions such that for some distribution $p$ and $\delta > 0$, $p_i = p$ for most $i$, and $||p_i - p||_1 \geq \delta$ for few outliers, where $||\cdot||_1$ is the $\ell_1$ distance between distributions. A sample from these set of distributions is an $m$-tuple $\overline{X} \stackrel{\text{def}}{=} (X_1, X_2, \ldots, X_m)$. Given *i.i.d.* samples from $\overline{p} \stackrel{\text{def}}{=} p_1, p_2, \ldots, p_m$, we would like to determine which are the outlier distributions.

Apart from the inherent theoretical interest to obtain good algorithms, universal outlier detection has several useful applications in diverse fields such as sensor networks [1], fraud detection [2], visual search in humans and animals [3], and target tracking [4].

We first relate outlier detection to the problem of closeness testing: given two sequences $X^{n_1}, Y^{n_2}$, generated independently from unknown distributions $p$ and $q$, closeness test decides if $p = q$ or $p \neq q$. Closeness testing is also known as *two-sample problem* or *homogeneity testing* [5].

For both the problems, similar to previous works, we focus on distributions $p_i$'s being discrete distributions over the same $k$ symbols. Classical approaches for such a problem are to use tools such as *likelihood ratio test (LRT)* or *generalized likelihood ratio test (GLRT)* [6]. These tools are typically studied in the asymptotic regime where the number of samples tend to infinity or more specifically number of samples is $\gg k$. Here in the asymptotic regime the objective is to show that the algorithms achieve the optimal error exponent. While the asymptotic approach characterizes the performance as the number of samples increases and often lends itself to elegant and simplistic tests, in many applications $k$ is large and the number of samples is $< k$ and the asymptotics do not to kick in. Owing to this reason, in the recent years many researchers have focused on algorithms for the non-asymptotic regime where the number of samples is limited and $< k$. Here rather than the error exponent, the objective is to design algorithms with good sample complexities [7]. Statistical problems such as *probability estimation*, closeness testing, and *classification* have received wide attention in this regime in the recent years see, [7–9] (and references there in). For outlier detection, we develop the first algorithm whose sample complexity is sublinear in $k$, *i.e.,* our outlier detector need not even see most of the symbols from most of the distributions.

## II. Related work

### A. Universal outlier detection

Universal outlier detection was first studied by [10, 11]. They considered two scenarios depending on whether $p$ is known or unknown and proposed a universal test motivated by GLRT. They studied error exponents and in particular whether the error probability decays exponentially in the number of samples, also called as *exponential consistency*. If $p$ is known, they proved that their test achieves the optimal error exponent. If $p$ is unknown, they showed that their test is universally exponentially consistent and moreover if $m \to \infty$, the error exponent converges to the optimal error exponent. These works assumed that the number of outliers were known. The assumption on number of outliers were removed in [12], where they showed that as long as the number of outliers is $> 0$, their test achieves universal exponential consistency and furthermore the assumption that number of outliers is $> 0$ is critical. In our work we focus on the case when $p$ is unknown and the number of outliers is unknown beforehand.

As stated before, instead of studying the error exponent in the asymptotic regime, we consider the problem in the sample-limited regime where the number of samples $< k$.

### B. Closeness testing

We now discuss recent results on closeness testing in the non-asymptotic regime. We refer readers to [5] (and references there in) for results in the asymptotic regime. Over the last decade, numerous researchers have studied closeness testing when both the sequences have same length *i.e.,* $n_1 = n_2$. We extend the sublinear closeness tests to handle sequences of unequal lengths and call it generalized closeness testing. When $n_1 = n_2$, [7] showed that closeness testing in $\ell_1$ distance requires sub-linear number of samples. They derived an algorithm that distinguishes pairs of identical distributions from pairs with $\ell_1$ distance $\geq \delta$ with error probability $\epsilon$ using $n_1 = n_2 = \mathcal{O}\left(\frac{k^{2/3} \log k}{\delta^4} \log \frac{1}{\epsilon}\right)$ samples. They also showed a lower bound of $\Omega\left(\frac{k^{2/3}}{\delta^{2/3}}\right)$ samples for error probability $1/3$. [13] removed the logarithmic factors and found the optimal-sample complexity upto constants when $n_1 = n_2$. Closeness testing restricted to a sub-class of distributions such as monotone, unimodal, or multi-modal were

considered in [14] and they found optimal sample complexities upto logarithmic factors.

[8, 15] showed that if the length of sequences are same *i.e.,* $n_1 = n_2 = n$ , then the sample complexity of closeness testing is similar to that of *classification*: where given two training sequences $X^n$, $Y^n$ from two unknown distributions and a test sequence $Z^n$, one asks which of the two underlying distributions generated $Z^n$. They constructed tests with sample complexity $\mathcal{O}(n_*^{3/2})$, where $n_*$ is the number of samples required by the optimal *label-invariant* test with prior information about distributions. They also showed that no test can achieve a sample complexity of $\mathcal{O}(n_*^{7/6})$ universally over all pairs of distributions.

[9] considered classification when all the probabilities are $\Theta\left(\frac{1}{k}\right)$, and showed that if the $\ell_1$ distance between the two underlying distributions is $> 0$, then comparing $\ell_2$ distance between the empirical frequencies of $X^n$ and $Z^n$ to those of $Y^n$ and $Z^n$ results in error probability strictly less than $\frac{1}{2}$ with $n = \mathcal{O}(\sqrt{k})$ samples.

## III. Notation and mathematical model

Without loss of generality we assume each $p_i$ is distributed over $[k] \stackrel{\text{def}}{=} 1, 2, \ldots k$ and the probability of symbol $j$ under distribution $p$ is denoted by $p(j)$. For outlier detection, the set of all distributions is denoted by $\overline{p} \stackrel{\text{def}}{=} p_1, p_2, \ldots, p_m$. We use $\mu(j)$ to denote the multiplicity, the number of occurances of symbol $j$. The empirical estimate of a symbol is denoted by $\hat{p}(j)$ and is $\hat{p}(j) = \frac{\mu(j)}{n}$. We use $||p-q||_1, ||p-q||_2, ||p-q||_\infty$ to denote $\ell_1$, $\ell_2$ and $\ell_\infty$ distance between $p$ and $q$ respectively. $T$ denotes closeness tests, and $D$ denotes outlier detector. $\mathbb{E}[\cdot]$ and $\mathrm{Var}(\cdot)$ denote the expectation and variance respectively. We use $X^n$ to denote a sequence of length $n$ and $X^*$ to denote sequences of all lengths. If we have multiple collection of distributions, we use $\overline{p}^1, \overline{p}^2 \ldots$ to differentiate between them. We use $\mathrm{poi}(\lambda)$ to denote a Poisson random variable with mean $\lambda$.

We now define closeness testing and outlier detection. From now on unless specified, all distributions are over $[k]$.

**Closeness testing:** Let $\mathcal{P}_\delta(k) \stackrel{\text{def}}{=} \{(p,q) : p = q \text{ or } ||p-q||_1 \geq \delta\}$, *i.e.,* $\mathcal{P}_\delta(k)$ contains pairs of distributions over $[k]$ that are same or have $\ell_1$ distance $\geq \delta$. For an underlying pair $(p,q) \in \mathcal{P}_\delta$, let $X^{n_1} \sim p$, $Y^{n_2} \sim q$ be the two test sequences. A *closeness test* is a mapping $T : [k]^* \times [k]^* \to \{\texttt{same}, \texttt{diff}\}$, where $T(x^*, y^*)$ indicates whether $x^*$ and $y^*$ are believed to be generated by the same or by different distributions. The error probability of $T$ is

$$\max_{(p,q)\in\mathcal{P}_\delta(k)} \begin{cases} \Pr\left(T(X^{n_1}, Y^{n_2}) = \texttt{same}\right) & \text{if } p \neq q, \\ \Pr\left(T(X^{n_1}, Y^{n_2}) = \texttt{diff}\right) & \text{else.} \end{cases}$$

**Outlier detection:** Let *typical* distribution be the one that is same for most indices in $[m]$. For $\overline{p}$, let $p$ be the typical distribution and $S(\overline{p})$ be the indices of the outlier distributions. Let $\overline{\mathcal{P}}_\delta(m,k,b)$ be

$$\{\overline{p} : \exists S, p \text{ s.t. } |S| \leq b, p_i = p \text{ for } i \notin S \ \&\ ||p - p_i||_1 \geq \delta \text{ else}\}.$$

Note that for the typical distribution to be unique, we need the number of outliers is $|S(\overline{p})| \leq b < m/2$. An outlier detector is a mapping $D : \left([k]^m\right)^* \to 2^{[m]}$, that observes $m$-tuple samples and outputs the set indices of the outlier distributions. The error probability of a outlier detector $D$ is

$$\max_{\overline{p}\in\overline{\mathcal{P}}_\delta(m,k,b)} \Pr\left(D(\overline{X}^n) \neq S(\overline{p})\right).$$

Similar to [7] and subsequent works on sublinear algorithms for property testing, we assume that our algorithm knows $\delta$. Note that without this assumption, for the error probability to go to 0, one needs to know the exact number of outliers [12].

## IV. Results

A simple result from [16] states that if we repeat an algorithm with new set of samples each time and take the majority output, then the error probability decreases.

*Lemma 1 ([16]):* Consider an algorithm that outputs correctly with probability $\geq 2/3$. Repeat it $12 \log \frac{1}{\epsilon}$ times independently and output the majority. The new algorithm has error $\leq \epsilon$.

Therefore, we restrict our attention to error probability $1/3$. The sample complexity for error $\epsilon$ is $12 \log \frac{1}{\epsilon}$ times the sample complexity for error $1/3$.

### A. Upper bounds

We first propose a closeness test for unequal length samples.

*Theorem 2:* Let $n_1 \geq n_2$, $n_2\sqrt{n_1} \geq 512\frac{k \log^{3/2} k}{\delta^3}$, and $n_2 \geq 128\frac{\sqrt{k \log k}}{\delta^2}$. If the number of samples from $p$ is $\geq 4n_1$ and the number of samples from $q$ is $\geq 2n_2$, then for all $(p,q) \in \mathcal{P}_\delta(k)$ Closeness test has error $\leq 1/4$.

We then use closeness test as a sub-routine for outlier detection and prove the following. The constant 12 in the corollary is for convenience. In general, our methods work for $|S(\overline{p})| < m/2$.

*Corollary 3:* If

$$n \geq \max\left(\frac{k^{2/3}b^{1/3}\log k}{\delta^2 m^{1/3}}, \frac{\sqrt{k \log k}}{\delta^2}\right) 6000 \log 4m,$$

then $\forall \overline{p} \in \overline{\mathcal{P}}_\delta(k,m,b)$ such that $|S(\overline{p})| \leq b \leq m/12$, Outlier detector with $\geq n$ samples has error $\leq 1/3$.

### B. Lower bounds

We show the following lower bounds for closeness testing and outlier detection.

*Theorem 4:* Let $n_1 \geq n_2$. There is a constant $c$ such that if $n_2 \leq c \cdot \frac{\sqrt{k}}{\delta^2}$ or $n_1\sqrt{n_2} \leq c \cdot \frac{k}{\delta^2}$, then for any closeness test has error $\geq 11/24$ for some $(p,q) \in \mathcal{P}_\delta(k)$.

By a simple reduction, we relate closeness testing to outlier detection and show the following lower bound.

*Corollary 5:* There is a constant $c$ such that if the number of samples $n \leq c \cdot \max\left(\frac{\sqrt{k}}{\delta^2}, \left(\frac{k}{m\delta^2}\right)^{2/3}\right)$, then every outlier detector makes an error of $\geq 11/24$ for some $\overline{p} \in \overline{\mathcal{P}}_\delta(k,m,b)$.

## V. Ancillary results

### A. $k = 2$

We are interested in the regime when the alphabet size $k$ is large. However, as a toy example we first consider the case when $k = 2$, *i.e.,* the underlying distributions are Bernoulli random variables $B(\cdot)$. More precisely, most distributions are

Bernoulli random variables with same $p$ and and the remaining distributions have parameters at least $\delta/2$ away from $p$. We outline a proof to show that the sample complexity for this problem is $\Theta\big(\frac{\log m}{\delta^2}\big)$.

**Upper bound:** A simple application of Chernoff bound and union bound shows that for some constant $c$ if $n \geq c \cdot \frac{\log m}{\delta^2}$, then with probability $\geq 2/3$, for all $i$, the empirical estimate $\hat{p}_i$ satisfies $|\hat{p}_i - p_i| < \delta/8$. Therefore the empirical estimates of typical distributions are $< \delta/4$ from each other. Moreover, the estimates from each outlier is at a distance $> \delta/4$ from them. Therefore if we cluster points together such that maximum distance within each cluster is $\leq \delta/4$, then the $S(\overline{p})$ would be the set of indices that are not in the largest cluster. A simple Single linkage [17] algorithm can be used for clustering.

**Lower bound:** We show that no test can detect all possible outliers with error $\leq 1/2$. Let $\delta \leq 1/10$. Consider the set of collection of distributions $\overline{p}^1 = (p_1^1, p_2^1, \ldots p_m^1), \overline{p}^2 = (p_1^2, p_2^2, \ldots p_m^2), \ldots \overline{p}^m = (p_1^m, p_2^m, \ldots p_m^m)$ where for $j \neq i$, $p_j^i = B(1/2)$ and $p_i^i = B(1/2 + \delta/2)$. If a test can detect the outlier using a certain number of samples, then given samples from one of $\overline{p}^1, \overline{p}^2, \ldots, \overline{p}^m$, it find the underlying $\overline{p}$. We show that no algorithm can differentiate between $\overline{p}^1, \overline{p}^2, \ldots, \overline{p}^m$ using $o\big(\frac{\log m}{\delta^2}\big)$ samples thus showing a lower bound on outlier detection. thus showing that there is no test By triangle inequality, The $\ell_1$ distance between any two distributions is $\geq \delta$. KL divergence between any two distributions is $\leq 4\delta^2$. Therefore by Fano's inequality [6] for any outlier detector, the error probability with $n$ samples is $\geq 1 - \frac{4n\delta^2 + \log 2}{\log m}$. For error probability to be small, the required number of samples is $\Omega\big(\frac{\log m}{\delta^2}\big)$.

While we have found the optimal sample complexity for $k = 2$, the same result does not hold for general $k$. A simple extension of the above result yields an upper bound of $\mathcal{O}(\frac{k \log m}{\delta^2})$ samples. However, as we show later a suitable choice of test yields a much better sample complexity.

### B. Poisson sampling

When a distribution $p$ is sampled $n$ times, the symbol multiplicities are mutually dependent, for example, they add to $n$. A standard approach to overcoming the dependence, *e.g.,* [16], samples the distribution a random number of times: $\mathrm{poi}(n)$, the Poisson distribution with mean $n$. Some useful properties of Poisson sampling include: (i) A symbol of probability $p$ appears $\mathrm{poi}(np)$ times. (ii) The numbers of times different symbols appear are independent of each other. (iii) For any fixed $n_0$, conditioned on the length $\mathrm{poi}(n) \geq n_0$, the distribution of the first $n_0$ elements is identical to sampling $p$ *i.i.d.* exactly $n_0$ times. We use $\mathrm{poi}(n)$ to denote a Poisson random variable with mean $n$.

For closeness testing, we assume that the number of samples from distributions to be $\mathrm{poi}(2n_1)$ and $\mathrm{poi}(n_2)$ respectively. We simulate $\mathrm{poi}(2n_1)$ samples from $4n_1$ samples by taking first $\mathrm{poi}(2n_1)$ samples from $4n_1$ samples. Similarly we simulate $\mathrm{poi}(n_2)$ from $2n_2$ samples. An additional error occurs if $\mathrm{poi}(2n_1) \geq 4n_1$ or $\mathrm{poi}(n_2) \geq 2n_2$. We relate errors of closeness tests with $\mathrm{poi}(2n_1), \mathrm{poi}(n_2)$ samples to $4n_1, 2n_2$ samples in the next lemma and it follows from tail bounds on Poisson random variables.

*Lemma 6:* Consider a closeness test with error $\leq 1/8$ for $\mathrm{poi}(2n_1), \mathrm{poi}(n_2)$ samples. Then there is an algorithm that has error $\leq 1/8 + 2e^{-\min(2n_1, n_2)/4}$ with $4n_1, 2n_2$ samples.

## VI. Sub-linear closeness test

Algorithm Closeness test
Let $n_1 \geq n_2$, $t = \log k + 2$, and $c = \frac{32t^2}{\delta^2}$
**Input**: $X^{\mathrm{poi}(2n_1)}$, $Y^{\mathrm{poi}(n_2)}$, $\delta$, and $k$
**Output**: `same` or `diff`

1) Use first $\mathrm{poi}(n_1)$ samples of $X$ to divide the symbols into sets: $B = \{i : \frac{c}{n_2} \leq \hat{p}(i)$ or $\frac{c}{n_2} \leq \hat{q}(i)\}$, $M = \{i : \frac{c}{n_1} \leq \hat{p}(i) < \frac{c}{n_2}\}$, and $S = \{i : \hat{p}(i) < \frac{c}{n_1}\}$
2) Divide $M$ into $M_j$ for $0 \leq j \leq \log\lceil \frac{n_1}{n_2} \rceil$ such that $i \in M_j$ if $\hat{p}(i) \in \left[\frac{c2^j}{n_1}, \frac{c2^{j+1}}{n_1}\right)$
3) Discard these samples and use the remaining samples to test:
   - `C1`: $\sum_{i \in B} |\hat{p}(i) - \hat{q}(i)| \leq \frac{\delta}{2t}$
   - `C2`: $\forall j$, $L_2(X^{\mathrm{poi}(n_1)}, Y^{\mathrm{poi}(n_2)}, M_j) \leq \frac{\delta^2 c 2^j}{2n_1}$
   - `C3`: $L_2(X^{\mathrm{poi}(n_1)}, Y^{\mathrm{poi}(n_2)}, S) \leq \frac{\delta^2}{2k}$
4) If `C1`, `C2`, and `C3` are satisfied output `same`, else `diff`

### A. Outline

Without loss of generality we assume $n_1 \geq n_2$. One of the natural tests would be to estimate the empirical $\ell_1$ distance $||\hat{p} - \hat{q}||_1$ and check if it is larger than $\delta/2$ or not. However if probabilities are smaller than $1/n_2$, then such a test has large variance and would not work. So we divide the probabilities into three sets and conduct different tests on each of the sets.

Closeness test uses $(\mathrm{poi}(2n_1), \mathrm{poi}(n_2))$ samples. Using the first $\mathrm{poi}(n_1)$ samples of $X$, it first divides the symbols in to big $B$, medium $M$, and small $S$ depending on the values of $\hat{p}(i)$. We further divide $M$ into $M_j$s such that $\hat{p}(i)$ within each $M_j$ differ at most by a factor of 2. To preserve independence, we discard these samples. Note that there are at most $\log n_1 + 2$ sets totally.

If $\sum_i |p(i) - q(i)| \geq \delta$, then for at least one of these sets $A$, $\sum_{i \in A} |p(i) - q(i)| \geq \frac{\delta}{\log n_1 + 2}$. Hence, using the remaining samples, we test on each of the sets $B, M_j, S$.

Since $B$ contains symbols with high empirical probabilities, symbols in it are likely to have high probabilities and therefore we can use empirical $\ell_1$ test.

Similary, symbols in sets $M$ and $S$ are likely to have small probabilities. We show that if the symbols have small probabilities, then $\ell_2$ distance can be estimated efficiently. The estimate accuracy improves if the symbols have similar probabilities. Therefore, we using first $\mathrm{poi}(n_1)$ samples from $X$ we further divide $M$ into $M_j$s to ensure better estimation. Note that $M$ plays an useful role only if *i.e.,* $n_1 \gg n_2$.

For $M_j$s and $S$ we use the $\ell_2$ test statistic (an estimate for $\ell_2$ distance), which is defined as follows:

$$L_2(X^{n_1}, Y^{n_2}, A) \stackrel{\mathrm{def}}{=} \sum_{i \in A} (\hat{p}(i) - \hat{q}(i))^2 - \frac{\hat{p}(i)}{n_1} - \frac{\hat{q}(i)}{n_2},$$

where $\hat{p}(i) = \frac{\mu(i, X^{n_1})}{n_1}$ and $\hat{q}(i) = \frac{\mu(i, Y^{n_2})}{n_2}$.

To preserve a good read, most of our proof uses just Chebyshev's inequalities or Markov inequality, instead of stronger Chernoff-type bounds. We believe few factors of $\log k$ can be removed with stronger bounds. For simplicity of the proof the algorithm assumes that $k \geq n_1 \log^2 k$. If $n_1 \log^2 k \geq k$, then substituting $n_1$ in the equations by $\min(k \log^{-2} k, n_1)$ gives the algorithm.

### B. Proof of Theorem 2

We first bound the variances and expected values of few of the test statistics in CLOSENESS TEST.

*Lemma 7:* Let $\mu_1 \sim \text{poi}(n_1\alpha)$ and $\mu_2 \sim \text{poi}(n_2\beta)$. If $Z = \left(\frac{\mu_1}{n_1} - \frac{\mu_2}{n_2}\right)^2 - \frac{\mu_1}{n_1^2} - \frac{\mu_2}{{n_2}^2}$, then $\mathbb{E}[Z] = (\alpha - \beta)^2$ and

$$\text{Var}\,(Z) = 2\left(\frac{\alpha}{n_1} + \frac{\beta}{n_2}\right)^2 + 4\left(\frac{\alpha}{n_1} + \frac{\beta}{n_2}\right)(\alpha - \beta)^2.$$

*Proof:* The expectation follows from the fact that $\mu_1, \mu_2$ are independent random variables with means $n_1\alpha$ and $n_2\beta$ and have $\mathbb{E}[\mu_1^2] = {n_1}^2\alpha^2 + n_1\alpha$ and $\mathbb{E}[\mu_2^2] = {n_2}^2\beta^2 + n_2\beta$. The variance calculation uses third and fourth moments of Poisson random variables. Proof is deferred to the full paper. ∎

Next, we show similar results for $\ell_1$-test statistic.

*Lemma 8:* Given $\text{poi}(n_1)$ samples from $p$ and $\text{poi}(n_2)$ samples from $q$, for every set $A$

$$0 \leq \mathbb{E}\big[\sum_{i\in A} |\hat{p}(i) - \hat{q}(i)| - |p(i) - q(i)|\big] \leq \sqrt{\frac{|A|}{n_1} + \frac{|A|}{n_2}},$$

and

$$\text{Var}(\sum_{i\in A} |\hat{p}(i) - \hat{q}(i)|) \leq \frac{1}{n_1} + \frac{1}{n_2}.$$

*Proof:* The lower bound on the expectation follows from Jensen's inequality and the fact that $|\cdot|$ is a convex function. For the upper bound, observe that

$$\begin{aligned}
&\left(\sum_{i\in A} \mathbb{E}\left[|\hat{p}(i) - \hat{q}(i)| - |p(i) - q(i)|\right]\right)^2 \\
&\overset{(a)}{\leq} |A| \sum_{i\in A} \mathbb{E}\left[(\hat{p}(i) - \hat{q}(i))^2 - (p(i) - q(i))^2\right] \\
&\overset{(b)}{\leq} |A| \sum_{i\in A} \left(\frac{p(i)}{n_1} + \frac{q(i)}{n_2}\right) \leq |A|\left(\frac{1}{n_1} + \frac{1}{n_2}\right).
\end{aligned}$$

$(a)$ follows from the Cauchy-Schwartz inequality and the lower bound on expectation. Lemma 7 implies $(b)$. For the variance, since the multiplicities are independent

$$\text{Var}(\sum_{i\in A} |\hat{p}(i) - \hat{q}(i)|) = \sum_{i\in A} \text{Var}(|\hat{p}(i) - \hat{q}(i)|).$$

A calculation similar to that of the upper bound for expectation can be used to show that for each $i$, $\text{Var}(|\hat{p}(i) - \hat{q}(i)|) = \frac{p(i)}{n_1} + \frac{q(i)}{n_2}$. Summing over the symbols yields the result. ∎

The last auxiliary lemma helps us bound $p(i)$ in terms of $\hat{p}(i)$ within sets $S$ or $M_j$. It is a simple application of Chernoff-bound and we omit the proof to conserve space.

*Lemma 9:* Let $t = \log k + 2$. Given $\text{poi}(n)$ samples from $p$, let $A$ be the set of symbols such that $\hat{p}(i) \leq c'/n$. Then

$$\Pr\left(\max_{i\in A} p(i) \geq \frac{4c' + 8t}{n}\right) \leq \frac{1}{16t}.$$

We now have all the tools to prove Theorem 2. We break it into two lemmas, we first show that if $p = q$, then the algorithm returns same with probability $\geq 7/8$.

*Lemma 10:* Let $t = \log k + 2$, $c = 32t^2\delta^{-2}$, and $n_1 \geq n_2$. If $p = q$, $n_2\sqrt{n_1} \geq 64\frac{k\sqrt{tc}}{\delta^2}$, and $n_2 \geq 64\frac{\sqrt{kt}}{\delta^2}$, then CLOSENESS TEST returns same with probability $\geq 7/8$.

*Proof:* Let $\texttt{C2}_j$ denote the condition $\texttt{C2}$ restricted to $M_j$. As stated before that proof assumes $n_1 \log^2 k \leq k$, otherwise replacing $n_1$ by $k \log^{-2} k$, yields the results. Observe that there are at most $2 + \log\lceil \frac{n_1}{n_2} \rceil \leq 2 + \log k = t$ sets, and hence at most $t$ conditions. We show that each condition fails with probability $\leq 1/(8t)$, thus showing that the total error probability $\leq 1/8$.

$\texttt{C1}$: Notice that $|B| \leq \frac{n_2}{c}$. Therefore by Lemma 8 and Chebyshev's inequality with probability $\geq 1 - 1/(8t)$,

$$\sum_{i\in B} |\hat{p}(i) - \hat{q}(i)| \leq \sqrt{\frac{2|B|}{n_2}} + \sqrt{\frac{16t}{n_2}} \leq \sqrt{\frac{2}{c}} + \sqrt{\frac{16t}{n_2}} \leq \frac{\delta}{2t}.$$

The last inequality follows from using the fact that $n_2 \gg \frac{t^3}{\delta^2}$ and substituting the value of $c = \frac{32t^2}{\delta^2}$.

$\texttt{C2}_j$: Note that

$$\begin{aligned}
L_2(X^{\text{poi}(n_1)}, Y^{\text{poi}(n_2)}, M_j) &\overset{(a)}{\leq} \frac{\sqrt{128t}}{n_2}\sqrt{\max_{i\in M_j} p(i)} \\
&\overset{(b)}{\leq} \frac{\sqrt{128t}}{n_2}\sqrt{\frac{4c2^{j+1} + 8t}{n_1}} \overset{(c)}{\leq} \frac{\delta^2 c 2^j}{2n_1}.
\end{aligned}$$

By Lemma 7 and Chebyshev's inequality, with probability $1 - 1/(16t)$ $(a)$ follows. By Lemma 9, with probability $\geq 1 - 1/(16t)$ $(b)$ follows. $(c)$ follows from the fact that $n_2 \geq \frac{64\sqrt{n_1}\sqrt{t}}{\delta^2\sqrt{c}}$. By the union bound, the total error is $\leq 1/(8t)$.

$\texttt{C3}$: Similar to the previous step, by Lemmas 7 and 9, with probability $\geq 1 - 1/(8t)$,

$$\begin{aligned}
L_2(X^{\text{poi}(n_1)}, Y^{\text{poi}(n_2)}, S) &\leq \frac{\sqrt{128t}}{n_2}\sqrt{\max_{i\in S} p(i)} \\
&\leq \frac{\sqrt{128t}}{n_2}\sqrt{\frac{4c + 8t}{n_1}} \leq \frac{\delta^2}{2k}.
\end{aligned}$$

The last inequality follows from the fact that $n_2\sqrt{n_1} \geq \frac{64k\sqrt{tc}}{\delta^2}$. The two conditions on sample complexity are $n_2 \geq \frac{64\sqrt{n_1}\sqrt{t}}{\delta^2\sqrt{c}}$ and $n_2\sqrt{n_1} \geq \frac{64k\sqrt{tc}}{\delta^2}$. It can be shown that under the assumption that $n_1 \log^2 k \leq k$, this is equivalent to the conditions $n_2 \geq \frac{64\sqrt{tk}}{\delta^2}$ and $n_2\sqrt{n_1} \geq \frac{64k\sqrt{tc}}{\delta^2}$. ∎

Similarly it can be shown that if $||p - q|| \geq \delta$, then the CLOSENESS TEST returns $\texttt{diff}$ with probability $\geq 7/8$.

*Lemma 11:* Let $t = \log k + 2$, $c = 32t^2\delta^{-2}$, and $n_1 \geq n_2$. If $||p - q||_1 \geq \delta$, $n_2\sqrt{n_1} \geq 64\frac{k\sqrt{c}}{\delta^2}$, and $n_2 \geq 64\frac{\sqrt{k}}{\delta^2}$, then CLOSENESS TEST returns diff with probability $\geq 7/8$.

The proof is similar to that of Lemma 10 and is omitted to conserve space. Theorem 2 follows directly from Lemmas 6, 10, and 11.

## VII. Sublinear outlier detector

Our outlier detector is a simple extension of Closeness test. Since there are at most $b \le m/12$ outliers, if we choose $\lfloor \frac{m}{12b} \rfloor$ indices at random, then with probability $\ge 1 - 1/12$ all of them are from the typical distribution $p$ and we can combine samples from these indices to obtain $n\lfloor \frac{m}{12b} \rfloor$ samples from $p$.

We run the closeness test described in the previous section between this set of samples from $p$ and each of the remaining indices repeatedly (as in Lemma 1) to get error probability $\le 1/4m$ for each index. Probability that our outlier detector fails is the sum of probabilities that we chose an index from $S(\overline{p})$ in the first step or closeness tests in one of the $m$ coordinates fails. By the union bound it is $\le 1/12 + m/4m = 1/3$.

The sample complexity follows from substituting $n_1 = n\lfloor \frac{m}{12b} \rfloor$ and $n_2 = n$ in Theorem 2. Note that for each coordinate, we have to repeat the closeness test $12 \log 4m$ times to get the error probability to $1/4m$, and hence we get an addition factor of $12 \log 4m$ in sample complexity.

## VIII. Lower bounds

### A. Proof of Theorem 4

Without loss of generality let $n_1 \ge n_2$. We first show that there is some constant $c$ such that, if $n_2 = c \cdot \sqrt{k}/\delta^2)$, then there is no closeness-test that has error $\le 1/3$ for all $(p, q) \in \mathcal{P}_\delta(k)$. Suppose $n_1 = \infty$. This condition is equivalent to knowing the distribution corresponding to $n_1$, *i.e.,* $p$. The problem reduces to finding if $Y^{n_2}$ is generated by $p$ or a distribution $\delta$-away from $p$. This problem has been well studied as *identity testing*, and a lower bound of $\Omega(\sqrt{k}/\delta^2)$ is known [18]. Therefore unless $n_2 \ge \Omega(\sqrt{k}/\delta^2)$, there is no closeness test with error $\le 1/3$ for all $(p, q) \in \mathcal{P}_\delta(k)$.

We now show that if $n_1\sqrt{n_2} \le \mathcal{O}\left(\frac{k}{\delta^2}\right)$, then are pairs of distributions that cannot be differentiated by any test. The proof of this lower bound is similar to that of [13]. Let $m^+(a, b) = \sum_{i=1}^k (n_1 p(i))^a (n_2 p(i))^b$ and $m^-(a, b) = \sum_{i=1}^k (n_1 p(i))^a (n_2 q(i))^b$. The second part of the proof uses the following variant of a result from [19].

*Lemma 12:* Without loss of generality let $n_1 \ge n_2$. If $\max(p(i), q(i)) \le \frac{1}{1000 n_1}$, and

$$\sum_{a,b:a+b\ge 2} \frac{|m^+(a, b) - m^-(a, b)|}{\lfloor \frac{a}{2}! \rfloor \lfloor \frac{b}{2} \rfloor ! \sqrt{1 + \max(m^+(a, b), m^-(a, b))}} < \frac{1}{360},$$

then there is no closeness test that differentiates between the two with $2n_1, 2n_2$ samples with error $\le 11/24$.

We now construct distributions $p$ and $q$ of support $k + 1000n_1 \le 1001k$. Since we are interested in the order, the constant 1001 does not affect our calculation. Let $p(i) = q(i) = \frac{1-\delta}{1000 n_1}$ for $1 \le i \le 1000n_1$, let $p(i) = \frac{2\delta}{k}$ for $1000n_1 + 1 \le i \le 1000n_1 + \frac{k}{2}$ and $q(i) = \frac{2\delta}{k}$ for $\frac{k}{2} + 1000n_1 + 1 \le i \le 1000n_1 + k$.

First note that if $\min(a, b) = 0$, then $m^+(a, b) = m^-(a, b)$ as the distributions are permutations of each other. For $\min(a, b) \ge 1$, a simple calculation shows that $m^+(a, b) - m^-(a, b) = k n_1^a {n_2}^b \left(\frac{2\delta}{k}\right)^{a+b}$ and $m^-(a, b) = 1000 n_1 \left(\frac{(1-\delta)^{a+b} n_1^a {n_2}^b}{(1000 n_1)^{a+b}}\right)$.

Since $\sum_{a,b:a+b\ge 2} \frac{1}{\lfloor \frac{a}{2}! \rfloor \lfloor \frac{b}{2} \rfloor !}$ is a convergent series and the $\frac{|m^+(a,b) - m^-(a,b)|}{\sqrt{m^-(a,b)}}$ falls exponentially with $a, b$, it is sufficient to show that the first term $\frac{|m^+(1,1) - m^-(1,1)|}{\sqrt{m^-(1,1)}} \le \frac{1}{1000}$. Computing we get

$$\frac{|m^+(1, 1) - m^-(1, 1)|}{\sqrt{m^-(1, 1)}} \le 4 \frac{\sqrt{n_1 n_2}\delta^2 \sqrt{2000 n_1}}{k}.$$

Therefore there is a constant $c$ such that if $n_1\sqrt{n_2} \le c \cdot \frac{k}{\delta^2}$, then no closeness test can distinguish between $p$ and $q$.

Thus we have shown that there is a constant $c$ such that if $n_2 \le c \cdot \frac{\sqrt{k}}{\delta^2}$ or $n_1\sqrt{n_2} \le c \cdot \frac{k}{\delta^2}$, then for any closeness test has error $\ge 11/24$ for some $(p, q) \in \mathcal{P}_\delta(k)$.

### B. Proof of Corollary 5

The proof follows from a simple reduction. Suppose we are given that $p_2 = p_3 = \ldots = p_m = p$, then outlier detection reduces to test if $p_1 = p$ or $||p_1 - p|| \ge \delta$, which is the problem of closeness testing. The number of samples we have from $p$ is the sum of samples from all coordinates $2, 3, \ldots, m$, *i.e.,* $(m-1)n$ and the number of samples from $p_1$ is $n$. We have shown a lower bound for closeness testing in Theorem 4. Substituting $n_1 = (m-1)n$ and $n_2 = n$ results in the corollary.

## IX. Acknowledgements

We thank Sirin Nitinawarat, Venkatadheeraj Pichapati, and Venugopal V. Veeravalli for helpful discussions.

## References

[1] J.-F. Chamberland and V. V. Veeravalli, "Wireless sensors in distributed detection applications," *Signal Processing Magazine, IEEE*, vol. 24, no. 3, pp. 16–25, 2007.
[2] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, pp. 235–249, 2002.
[3] N. K. Vaidhiyan, S. P. Arun, and R. Sundaresan, "Active sequential hypothesis testing with application to a visual search problem," in *Proc. of the ISIT*, 2012, pp. 2201–2205.
[4] L. D. Stone, *Theory of optimal search*. Academic Press New York, 1975.
[5] J. Unnikrishnan, "On optimal two sample homogeneity tests for finite alphabets," in *Proc. of the ISIT*, 2012, pp. 2027–2031.
[6] T. M. Cover and J. A. Thomas, *Elements of information theory (2. ed.)*. Wiley, 2006.
[7] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *Proc. of the 41st FOCS*, 2000, pp. 259–269.
[8] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan, "Competitive closeness testing," *Proc. of the 24th COLT*, vol. 19, pp. 47–68, 2011.
[9] B. G. Kelly, A. B. Wagner, T. Tularak, and P. Viswanath, "Classification of homogeneous data with large alphabets," *IEEE Trans. on Info. Theory*, vol. 59, no. 2, pp. 782–795, 2013.
[10] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," in *Proc. of the ISIT*, 2013, pp. 2666–2670.
[11] ——, "Universal outlier detection," *CoRR*, vol. abs/1302.4776, 2013.
[12] ——, "Universal multiple outlier hypothesis testing," in *2013 IEEE 5th Intl. Workshop on CAMSAP*, 2013, pp. 177–180.
[13] S. O. Chan, I. Diakonikolas, P. Valiant, and G. Valiant, "Optimal algorithms for testing closeness of discrete distributions," in *Proc. of the 25th Annual SODA*, 2014, pp. 1193–1203.
[14] C. Daskalakis, I. Diakonikolas, R. A. Servedio, G. Valiant, and P. Valiant, "Testing $k$-modal distributions: Optimal algorithms via reductions," in *Proc. of the 24th Annual SODA*, 2013, pp. 1833–1852.
[15] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. T. Suresh, "Competitive classification and closeness testing," in *Proc. of the 25th COLT*, 2012, pp. 22.1–22.18.
[16] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.
[17] R. Sibson, "Slink: An optimally efficient algorithm for the single-link cluster method," *The Computer Journal*, vol. 16, no. 1, pp. 30–34, 1973.
[18] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Trans. on Info. Theory*, vol. 54, no. 10, pp. 4750–4755, 2008.
[19] P. Valiant, "Testing symmetric properties of distributions," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1927–1968, December 2011.