JMLR: Workshop and Conference Proceedings vol 23 (2012) 22.1–22.18 25th Annual Conference on Learning Theory

Competitive Classification and Closeness Testing

Jayadev Acharya JAYADEV@UCSD.EDU
Hirakendu Das HDAS@UCSD.EDU
Ashkan Jafarpour ASHKAN@UCSD.EDU
Alon Orlitsky ALON@UCSD.EDU
Shengjun Pan S1PAN@UCSD.EDU
Ananda Suresh ASURESH@UCSD.EDU
University of California San Diego, La Jolla, CA 92093

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

We study the problems of *classification* and *closeness testing*. A *classifier* associates a test sequence with the one of two training sequences that was generated by the same distribution. A *closeness test* determines whether two sequences were generated by the same or by different distributions. For both problems all natural algorithms are *symmetric*—they make the same decision under all symbol relabelings. With no assumptions on the distributions' support size or relative distance, we construct a classifier and closeness test that require at most $\tilde{\mathcal{O}}(n^{3/2})$ samples to attain the n-sample accuracy of the best symmetric classifier or closeness test designed with knowledge of the underlying distributions. Both algorithms run in time linear in the number of samples. Conversely we also show that for any classifier or closeness test, there are distributions that require $\tilde{\Omega}(n^{7/6})$ samples to achieve the n-sample accuracy of the best symmetric algorithm that knows the underlying distributions.

Keywords: Classification, closeness testing, competitiveness

1. Background

Classification is one of the most studied problems in statistics and machine learning. In its simplest form, given two *training sequences* $\overline{X}$ and $\overline{Y}$, each generated by an unknown *i.i.d.* distribution, a *classifier* attempts to determine which of the two distributions generated a *test sequence* $\overline{Z}$.

Traditional classifiers include the *likelihood ratio test (LRT)*, the *generalized likelihood ratio test (GLRT)*, and the Chi-square test, as well as classifiers outlined in Ziv (1988); Gutman (1989). The standard analysis is in the asymptotic regime where the underlying alphabet size k is fixed and the sample size n, for simplicity assumed to be the same for $\overline{X}$, $\overline{Y}$, and $\overline{Z}$, tends to infinity.

However, in many applications, the sample size is small relative to the alphabet size. For example, the length of text documents is often much smaller than the size of the English vocabulary, hence the asymptotic analysis does not apply. Recent work has therefore focused on classification accuracy for general ranges of k and n.

If $n = o(\sqrt{k})$ and the distributions are close to uniform, the birthday problem suggests that every element in $\overline{X}$, $\overline{Y}$, and $\overline{Z}$, will appear just once, and classification is impossible. Kelly et al. (2010) considered the complementary range, $n = \Omega(k^\alpha)$ for $\alpha > \frac{1}{2}$. They showed that even for this range, the above techniques may not work, *e.g.,* Theorem 5 therein or the simpler Example 1 in Acharya et al. (2011), but proved that if all symbol probabilities are $\Theta\left(\frac{1}{k}\right)$, and the ℓ_1 distance

between the two underlying distributions is bounded away from 0, then comparing the ℓ_2 distance between the empirical frequencies of $\overline{X}$ and $\overline{Z}$ to those of $\overline{Y}$ and $\overline{Z}$ results in error probability strictly less than $\frac{1}{2}$. It may be worth observing that the ℓ_2-distance criterion corresponds to the classification performed by the simplest possible support-vector-machine with linear kernel and just two classes as defined *e.g.,* in Scholkopf and Smola (2001). However, the ℓ_2-distance test fails if the probabilities are not in the same range and the condition on ℓ_1 is relaxed.

Example 1 *Consider the distributions* $P = (0.5, 0.5 - \delta, \delta, 0)$ *and* $Q = (0.5, 0.5 - \delta, 0, \delta)$. *The* ℓ_2 *distance test cannot classify them with* $100/\delta$ *samples, while* GLRT *can classify them with error less than* e^{-25} *in* $100/\delta$ *samples.*

While showing that classification can be performed when the number of samples is sub-linear in the support size, the probability and distance assumptions substantially limit the class of distributions for which this result applies. To classify a wider range of distributions, we draw on the related problem of *closeness-testing* that asks whether two sequences $\overline{X}$ and $\overline{Y}$ were generated by the same or by different distributions. For a precise connection between the two problems, see Lemma 3.

Over the last decade, closeness testing was considered by a number of researchers. Batu et al. (2000) showed that closeness testing requires a sub-linear number of samples. They derived an algorithm that distinguishes pairs of identical distributions from pairs with ℓ_1 distance $\geq \delta$ with error probability ϵ using $n = \mathcal{O}\left(\frac{k^{2/3}\log k}{\delta^4}\log\frac{1}{\epsilon}\right)$ samples. They also constructed two distributions requiring at least $\Omega\left(\frac{k^{2/3}}{\delta^{2/3}}\right)$ samples for error probability $1/3$. Valiant (2008) showed that for $\delta_1 < \delta_2$, distinguishing between distribution pairs with ℓ_1 distance $\leq \delta_1$ and those where it is $\geq \delta_2$ requires between $k^{1-o(1)}$ and $\mathcal{O}(k)$ samples. Guha et al. (2009) derived $\mathcal{O}(k^{2/3})$ algorithms for testing closeness in any f-divergence, such as the Hellinger and Chi-square norms, and Valiant and Valiant (2011) considered linear estimators for ℓ_1 distance and KL divergence of distribution pairs. Biau and Gyorfi (2005) constructed tests for the closely related problem of testing homogeneity of two independent multivariate samples. Similar problems arising in property testing were considered in Batu et al. (2001); Raskhodnikova (2003).

All of the above algorithms require an a priori knowledge of both an upper bound on the support size k and of a lower bound on the distance δ between the two distributions. These requirements often limit the applicability of the results. Many popular and useful distributions, such as Poisson, Zipf, or the distribution of English words, have infinite or very large support, rendering the algorithms impractical, or the guarantees weak. And even when the support is finite and relatively small, in many applications we do not have an upper bound on it, and therefore cannot construct the algorithms. The same holds for the distance δ between the distributions.

Additionally, the support size k and the distance δ between the underlying distributions may not be very indicative of the sample complexity. The next example shows two natural pairs of distributions with essentially the same support size, yet the pair with the larger distance between the two distributions requires many more samples to classify.

Example 2 *Let* $P_1 = U\{1, \ldots, k\}$ *and* $Q_1 = U\{k+1, \ldots, 2k\}$, *while* $P_2 = U\{1, \ldots, k\}$ *and* $Q_2(0) = \delta$ *and* $Q_2(1) = \ldots = Q_2(k) = \frac{1-\delta}{k}$. *Observe that all distributions have essentially the same support size, and that* $\ell_1(P_1, Q_1) = 2 \gg 2\delta = \ell_1(P_2, Q_2)$. *Yet the distribution with the smaller* ℓ_1 *distance is easier to classify. By the birthday paradox, any symmetric test requires at least* $\sqrt{k}$ *samples to classify* P_1 *and* Q_1. *On the other hand, by considering just the most frequent element, the pair* (P_2, Q_2), *with the lower-*ℓ_1*-distance, can be classified using just* $\mathcal{O}(\frac{\log k}{\delta})$ *samples.*

For these reasons, Acharya et al. (2011) considered a competitive approach to closeness testing. Instead of analyzing the performance relative to the distributions' alphabet size or some distance between them, they considered the lowest error probability achieved by the best symmetric closeness test designed specifically for the given underlying distributions. They constructed a general closeness test T that is based on knowledge of just ϵ, and not the alphabet size or the distance between the distributions, and showed that for any distribution pair (P, Q), if any test uses n samples to distinguish this pair from all identical distribution pairs with error probability ϵ, then T achieves the same error probability using $\mathcal{O}\left(\frac{n^3}{\log(1/\epsilon)}\right)$ samples.

Adopting the competitive-optimality framework, we derive a general closeness test and a classifier that use $\tilde{\mathcal{O}}(n^{3/2})$ samples to achieve the best accuracy possible with n samples by algorithms designed with knowledge of the underlying distributions. The closeness test needs as a parameter an upper bound on the error probability, but the classifier is constructed without any parameters.

Conversely, we show that for *every* closeness test or classifier, there are distributions where these algorithms would require at least $\tilde{\Omega}(n^{7/6})$ samples to attain the same accuracy achieved with n samples by an algorithm that knows the underlying distributions.

2. Definitions

A *closeness test* is a mapping $T : \mathcal{A}^* \times \mathcal{A}^* \to \{\texttt{same}, \texttt{diff}\}$, where $T(\overline{x}, \overline{y})$ indicates whether $\overline{x}$ and $\overline{y}$ are believed to be generated by the same or by different distributions. T wrongly declares P and Q as `same` or `diff` based on n samples with probability

$$\bar{\mathcal{E}}^T_{P,Q}(n) \stackrel{\text{def}}{=} \begin{cases} \Pr\left(T(X^n, Y^n) = \texttt{same}\right) & \text{if } P \neq Q, \\ \Pr\left(T(X^n, Y^n) = \texttt{diff}\right) & \text{if } P = Q, \end{cases}$$

where $(X^n, Y^n) \sim (P, Q)$. The worst-case error of T for two distinct distributions $P \neq Q$ based on n samples is

$$\mathcal{E}^T_{P,Q}(n) \stackrel{\text{def}}{=} \max\left(\bar{\mathcal{E}}^T_{P,Q}(n), \max_R \{\bar{\mathcal{E}}^T_{R,R}(n)\}\right),$$

the highest probability that T declares (P, Q) same, or declares any identical *i.i.d.* distributions R different. For $P = Q$, we define $\mathcal{E}^T_{P,Q}(n) \stackrel{\text{def}}{=} \frac{1}{2}$.

Since no prior knowledge is assumed, any reasonable classifier or closeness test must be *symmetric*, *e.g.,* see Batu (2001); Acharya et al. (2011). In other words, its outcome should remain the same under any relabeling of the symbols. For example, if a closeness test declares two sequences aab and cbc to be generated by different distributions, it must reach the same conclusion for uut and gtg. Similarly if given training sequences aab and cbc, a classifier associates abd with aab, then given uut and gtg, the classifier should associate utz with uut. We therefore compare the performance of a classifier with that of the best symmetric classifier designed with knowledge of the underlying distributions.

Let $\mathcal{T}$ be the collection of symmetric closeness tests. For every distribution pair $P \neq Q$,

$$\mathcal{E}^t_{P,Q}(n) \stackrel{\text{def}}{=} \min_{T \in \mathcal{T}} \mathcal{E}^T_{P,Q}(n),$$

is the lowest error probability of any symmetric test, including those designed with prior knowledge of P and Q. Note that such a test has error probability at most $\mathcal{E}^t_{P,Q}$ for (P, Q) as well as all pairs of identical distributions R and R.

A classifier is a mapping $S : \mathcal{A}^* \times \mathcal{A}^* \times \mathcal{A}^* \to \{x, y\}$, where $S(\overline{x}, \overline{y}, \overline{z})$ indicates whether $\overline{z}$ is generated by the same distribution as $\overline{x}$ or $\overline{y}$. Its error probability is

$$\mathcal{E}^S_{P,Q}(n) \stackrel{\text{def}}{=} \begin{cases} \Pr\left(S(X^n, Y^n, Z^n) = \mathtt{y}\right) & \text{if } Z^n \sim X^n, \\ \Pr\left(S(X^n, Y^n, Z^n) = \mathtt{x}\right) & \text{if } Z^n \sim Y^n. \end{cases}$$

Let $\mathcal{S}$ be the collection of symmetric classifiers. For every distribution pair $P \neq Q$, let

$$\mathcal{E}^s_{P,Q}(n) \stackrel{\text{def}}{=} \min_{S \in \mathcal{S}} \mathcal{E}^S_{P,Q}(n),$$

be the lowest error achieved for P and Q by any symmetric classifier, where the classifier achieving this lowest error is typically designed with prior knowledge of P and Q.

A closeness test or classifier for n samples, can be used to construct closeness tests or classifiers for multiples of n samples with exponentially smaller error probability. Simply partition the larger sample into groups of n samples, use the original test on each group, and take a majority decision. This notion is quantified in Lemma 4.

Hence once the error falls below some fixed value $< \frac{1}{2}$, any error ϵ can be achieved with just $C \log \frac{1}{\epsilon}$ times the number of samples. It is therefore of interest to determine when the error falls below some fixed value, which for simplicity we take to be 10%. Let

$$N^T_{P,Q}(\epsilon) \stackrel{\text{def}}{=} \min\{n : \mathcal{E}^T_{P,Q}(n) \leq \epsilon\} \qquad \text{and} \qquad N^t_{P,Q}(\epsilon) \stackrel{\text{def}}{=} \min\{n : \mathcal{E}^t_{P,Q}(n) \leq \epsilon\}$$

be the smallest number of samples a test T, and the best test for (P, Q), need to achieve error probability ϵ. Similarly for classification

$$N^S_{P,Q}(\epsilon) \stackrel{\text{def}}{=} \min\{n : \mathcal{E}^S_{P,Q}(n) \leq \epsilon\} \qquad \text{and} \qquad N^s_{P,Q}(\epsilon) \stackrel{\text{def}}{=} \min\{n : \mathcal{E}^s_{P,Q}(n) \leq \epsilon\}.$$

Finally, we will typically use $\epsilon = 1/10$, hence will abbreviate

$$N^T_{P,Q} \stackrel{\text{def}}{=} N^T_{P,Q}(1/10).$$

Similarly for $N^t_{P,Q}(1/10)$, $N^S_{P,Q}(1/10)$ and $N^s_{P,Q}(1/10)$.

3. Results

In Section 4 we relate classification and closeness testing. In Lemma 3 we show that classification is strictly easier, but not much more so than closeness testing.

In Section 5 we construct a symmetric closeness test T described in Algorithm 1 and a symmetric classifier S described in Algorithm 2. In Theorem 8 we bound the competitive sample complexities of these algorithms, showing that for every (P, Q),

$$N^T_{P,Q} \leq \mathcal{O}({N^t_{P,Q}}^{1.5} \log N^t_{P,Q}) \qquad \text{and} \qquad N^S_{P,Q} \leq \mathcal{O}({N^s_{P,Q}}^{1.5} \log N^s_{P,Q}),$$

where the algorithms and implied constants do not depend on the alphabet size, the distributions, or anything else.

For any smaller error probability, these algorithms can be combined with a simple majority decision to derive closeness test T' and classifier S' with a similar competitive performance. In Corollary 9 we bound the sample complexity to achieve a lower error probability ϵ as

$$N^{T'}_{P,Q}(\epsilon) \leq \mathcal{O}\left({N^t_{P,Q}}^{1.5} \log N^t_{P,Q} \log \frac{1}{\epsilon}\right) \qquad \text{and} \qquad N^{S'}_{P,Q}(\epsilon) \leq \mathcal{O}\left({N^s_{P,Q}}^{1.5} \log N^s_{P,Q} \log \frac{1}{\epsilon}\right).$$

In Section 6 we prove lower bounds on competitive guarantees for sample complexity. In Theorem 11 we show that for every closeness test T and classifier S there are distribution pairs (P, Q) (which may differ for the two cases) such that

$$N^T_{P,Q} \geq \Omega\left(\frac{{N^t_{P,Q}}^{7/6}}{\log N^t_{P,Q}}\right) \qquad \text{and} \qquad N^S_{P,Q} \geq \Omega\left(\frac{{N^s_{P,Q}}^{7/6}}{\log N^s_{P,Q}}\right).$$

4. Preliminaries

Profiles

The *multiplicity* $\mu(x)$ of a symbol x in a sequence is the number of times it appears. The *profile* φ' of a sequence is the multiset of multiplicities of all symbols appearing in it Orlitsky et al. (2004b,a). For example, the sequence $ababcde$ has multiplicities $\mu(a) = \mu(b) = 2$, $\mu(c) = \mu(d) = \mu(e) = 1$, and profile $\{1, 1, 1, 2, 2\}$.

The *joint profile* Dhulipala and Orlitsky (2006); Acharya et al. (2010b) $\varphi(\overline{x}, \overline{y})$ of two sequences $\overline{x}$ and $\overline{y}$ is the multiset φ of pairs of multiplicities of all symbols appearing in at least one of them. For example, $\varphi(ababcde, babbdef) = \{(2, 3), (2, 1), (1, 1), (1, 1), (1, 0), (0, 1)\}$. The prevalence φ_μ (respectively, $\varphi_{(\mu,\mu')}$) of μ (respectively (μ, μ')) is the number of times it appears in the profile. For example, in the above, $\varphi_{(1,1)} = 2$. For symmetric closeness and classification tests under *i.i.d.* sampling, the joint profiles of the observed sequences are a sufficient statistic, hence we consider only tests that operate on the joint profile.

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.,* Mitzenmacher and Upfal (2005), samples the distribution a random number of times $\sim \mathrm{Poi(n)}$, the Poisson distribution with parameter n. Some useful properties of Poisson sampling include:

Fact 1 *If a distribution P is sampled i.i.d.* $\mathrm{Poi(n)}$ *times, then:*

- *A symbol of probability p appears* $\mathrm{Poi(np)}$ *times.*
- *The numbers of times different symbols appear are independent of each other.*
- *For any fixed n_0, conditioned on the length* $\mathrm{Poi(n)} \geq \mathrm{n}_0$*, the distribution of the first n_0 elements is identical to sampling P i.i.d. exactly n_0 times.* ■

Next we relate the closeness-test error with $\mathrm{Poi(2n)}$ samples to that of exactly n samples.

Lemma 2 *For all (P, Q) pairs,* $\mathcal{E}^t_{P,Q}(\mathrm{Poi(2n)}) \leq \mathcal{E}^{\mathrm{t}}_{\mathrm{P,Q}}(\mathrm{n}) + 2\mathrm{e}^{-\mathrm{n}/4}$.

Proof When a distribution is sampled $\mathrm{Poi}(2n)$ times, the simple Poisson-tail bound of Fact 12 in the Appendix shows that the sequence length is $< n$ with probability $\le e^{-n/4}$. Hence the probability that one of the sequences has $< n$ samples is $\le 2e^{-n/4}$. With the remaining probability, both samples have length $\ge n$, and Fact 1 along with a standard closeness test on n show that the error probability is at most $\mathcal{E}^t_{P,Q}(n) + 2e^{-n/4}$. ■

Profile probabilities

The probability of a profile φ' under a distributions P sampled $\mathrm{Poi}(n)$ times is the probability of observing a sequence with profile φ'. Similarly the probability of a joint profile φ is the probability of observing sequence pairs with joint profile φ.

$$\Pr(\varphi) = \sum_{\varphi(\overline{x},\overline{y})=\varphi} \Pr(\overline{x},\overline{y}).$$

Let $P_{1,2}$ denote the product distribution of P and Q. For $A \subset \mathcal{A}$, the *sub-profile* φ_A of a sequence as the set of non-zero joint-multiplicities of symbols in A. For example, for $A = \{a, c, g\}$, the sequence pair $ababcde$ and $babbdef$ has $\varphi_A = \{(2,1),(1,0)\}$. The probability of a sub-profile $\varphi_A = \{(\mu_1,\mu'_1),(\mu_2,\mu'_2),\ldots\}$ when P and Q are sampled $\mathrm{Poi}(n)$ times is

$$P_{1,2}(\varphi_A) = \frac{\sum_\sigma \prod_{i\in A} e^{-np_i - nq_i}(np_i)^{\mu_{\sigma_i}}(nq_i)^{\mu'_{\sigma_i}}}{N_d(\varphi_A)},$$

where the summation is over all symbol permutations, and

$$N_d(\varphi_A) \stackrel{\text{def}}{=} \prod_{\mu_1,\mu_2} \varphi_{\mu_1,\mu_2}!(\mu_1!\mu_2!)^{\varphi_{\mu_1,\mu_2}} \tag{1}$$

is related to the number of patterns of a profile φ_A. In the above example,

$$P_{1,2}(\varphi_A) = \frac{1}{2}n^3 e^{-n(p_a+p_c+p_g+q_a+q_c+q_g)}\left(p_a^2q_a(p_c+p_g) + p_c^2q_c(p_a+p_g) + p_g^2q_g(p_a+p_c)\right).$$

Due to the large number of permutations, $\Pr(\varphi_A)$ is hard to analyze. Various techniques to compute them are studied in Zhang (2005); Acharya et al. (2010a); Orlitsky et al. (2012). Clearly, when $A = \mathcal{A}$, we obtain the probability of φ. The equation is explained in Acharya et al. (2011). A related quantity that will be used in the closeness test is

$$N_s(\varphi_A) \stackrel{\text{def}}{=} \prod_{\mu''}\Big(\sum_{\mu+\mu'=\mu''} \varphi_{\mu,\mu'}\Big)!(\mu''!)^{(\sum_{\mu+\mu'=\mu''}\varphi_{\mu,\mu'})} \frac{1}{2^{\sum_{\mu,\mu'}\varphi_{\mu,\mu'}(\mu+\mu')}}. \tag{2}$$

Properties of closeness testing and classification

We first show that closeness testing and classification have similar sample complexities. If two distributions can be classified by n samples, then by few more samples they can be tested for closeness. Conversely, if they can be tested for closeness in n samples, by few more samples they can be classified. The proof is given in Appendix B.1.

Algorithm 1 Closeness Test $\mathrm{T}(\overline{x}, \overline{y})$

Parameters: Constants c_1 and c_2 determined later in the proofs irrespective of the distributions and n.
Input: Two sequences $\overline{x}$ and $\overline{y}$ of length $4c_1 n^{3/2} \log n$.
Output: `same` or `diff`

$n_1 \leftarrow \mathrm{Poi}(\mathrm{c_1 n^{3/2} \log n})$ $\quad \overline{x}_1 \leftarrow \overline{x}_1^{n_1}$
$n_2 \leftarrow \mathrm{Poi}(\mathrm{c_1 n^{3/2} \log n})$ $\quad \overline{x}_2 \leftarrow \overline{x}_{n_1+1}^{n_1+n_2}$
$n_3 \leftarrow \mathrm{Poi}(\mathrm{n_0} \stackrel{\mathrm{def}}{=} \mathrm{c_2 n \log^3 n})$ $\quad \overline{x}_3 \leftarrow \overline{x}_{n_1+n_2+1}^{n_1+n_2+n_3}$.

Repeat for $\overline{y}$.
$\mu_i(\overline{x}) \leftarrow$ multiplicity of symbol i in $\overline{x}$.
$A \leftarrow \{i \mid \mu_i(\overline{x}_1) + \mu_i(\overline{y}_1) \geq \frac{c_1\sqrt{n}}{\log n}\}$.

C1 $\leftarrow \sum_{i \in A} \frac{(\mu_i(\overline{x}_2) - \mu_i(\overline{y}_2))^2 - \mu_i(\overline{x}_2) - \mu_i(\overline{y}_2)}{\mu_i(\overline{x}_2) + \mu_i(\overline{y}_2) - 1} \geq \frac{c_1}{24}\sqrt{n}\log n$.

C2 $\leftarrow \frac{N_d(\varphi_{A^c})}{N_s(\varphi_{A^c})} \geq 34e^{120\log^3 n_0}$ for $\overline{x}_3$ and $\overline{y}_3$, where N_s and N_d are defined in Equations (1) and (2).

if $C1 = \mathtt{true} \vee C2 = \mathtt{true}$ **then**
  return `diff`
else
  return `same`
end if

Lemma 3 *For all* (P, Q) *and* $\epsilon < \frac{1}{2}$,

$$\frac{1}{\log(1/\epsilon)} \cdot N^t_{P,Q}\left(\epsilon \log \frac{1}{\epsilon}\right) \leq N^s_{P,Q}(\epsilon) \leq N^t_{P,Q}(\epsilon). \qquad \blacksquare$$

Next we prove that once the error falls below some number strictly smaller than 0.5, then lower error probabilities can be achieved by a simple majority decision. Using this lemma we will relate sample complexities for any error $\epsilon \leq 1/10$ to that of $1/10$.

Lemma 4 *For all* (P, Q), n, *and* m

$$\mathcal{E}^t_{P,Q}(nm) \leq \frac{1}{2}\left(2\sqrt{\mathcal{E}^t_{P,Q}(n)(1 - \mathcal{E}^t_{P,Q}(n))}\right)^m \text{ and } \mathcal{E}^s_{P,Q}(nm) \leq \frac{1}{2}\left(2\sqrt{\mathcal{E}^s_{P,Q}(n)(1 - \mathcal{E}^s_{P,Q}(n))}\right)^m.$$

Proof See Appendix B.2. $\blacksquare$

5. Algorithms for closeness testing and classification

Broadly speaking, we separately consider symbols with *high* probability ($\geq \frac{1}{n \log^2 n}$) and those *small* probability. We first construct a test with error at most $\sqrt{\epsilon}$ that uses either only the sub-profile of the high-probability elements or only the sub-profile of the low-probability elements.

For any distribution pair, if there is a test that uses only high probability elements to achieve error $\sqrt{\epsilon}$, then we show a test that has error at most ϵ when $cn^{3/2} \log n$ samples are provided for some constant c. The algorithm estimates the Chi-square distance between distributions from the multiplicities. If the distributions were different, then we show that the Chi-square distance must be at least a certain quantity and then show that given $cn^{3/2} \log n$ samples, the expected value of

Algorithm 2 Classifier $S(\overline{x}, \overline{y}, \overline{z})$

Parameters: A constant c_3 determined later in the proof irrespective of the distributions, and n.
Input: Three sequences $\overline{x}, \overline{y}$ and $\overline{z}$ of length $c_3 n^{1.5} \log n$.
Output: `x` or `y`.
if $T(\overline{x}, \overline{z}) =$ `same` **then**
    return `x`
else
    return `y`
end if

the estimate is much larger than its standard deviation, hence attaining small error probability. A similar argument works for identical distributions.

Similarly, for any distribution pair for which there is a test that uses only the small sub-profile we use a variation of the $N(\varphi)$ test proposed in Acharya et al. (2011). The algorithm relies on the observation that when all probabilities are small, then the number of *likely* joint profiles is *small*. Using this, we bound the probability of a sub-profile efficiently using a deterministic function of φ_A.

In the following section we provide a complete proof of the correctness of the algorithm and explain the steps in greater detail. In Lemma 3, we prove that closeness tests and classifiers are closely related. Hence, the classifier tests for closeness between $\overline{x}$, $\overline{z}$ and $\overline{y}$, $\overline{z}$ and outputs the sequence for which the output is `same`. The classifier given in Algorithm 2. Using Algorithm 2, we prove the latter part of Theorem 8. Next we relate the error probability of an optimal test to the Chi-Square distance between P and Q, and use this relation to bound $N^t_{P,Q}(\epsilon)$. For simplicity we denote the joint probability distribution of P and Q by $P_{1,2}$ and the joint probability of R and R by $P_{3,3}$.

Lemma 5 *If $\mathcal{E}^t_{P,Q}(\mathrm{Poi(n)}) \le \epsilon$, then for any set $A \subseteq \mathcal{A}$ at least one of the following holds:*

1. *For any possible sub-profile φ_{A^c}, either $P_{1,2}(\varphi_{A^c}) \le \sqrt{2\epsilon}$, or $P_{3,3}(\varphi_{A^c}) < \sqrt{2\epsilon}$ for all distributions R.*
2. $$\sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} \ge \frac{1}{2n} \log\left(\frac{1}{8\epsilon}\right).$$

Proof Since the $P_{1,2}$ (and therefore $P_{3,3}$) are unknown, the error $\mathcal{E}^t_{P,Q}(\mathrm{Poi(n)})$ is larger than that of the hypothesis testing problem with two equally likely hypotheses—one where the joint profile is generated by $P_{1,2}$ and the other by $P_{3,3}$. This error in turn decreases if we are told which multiplicities derive from A, and which from A^c. Therefore, identifying φ_A with the set of sequences of that sub-profile, we see that $\varphi_{A\cap B} = \varphi_A \cap \varphi_B$, we obtain

$$\begin{aligned}
\epsilon \ge P_e &= \frac{1}{2} \sum_{\varphi_A, \varphi_{A^c}} \min\left(P_{1,2}(\varphi_A \cap \varphi_{A^c}), P_{3,3}(\varphi_A \cap \varphi_{A^c})\right) \\
&\ge \frac{1}{2} \left(\sum_{\varphi_A} \min\left(P_{1,2}(\varphi_A), P_{3,3}(\varphi_A)\right)\right) \left(\sum_{\varphi_{A^c}} \min\left(P_{1,2}(\varphi_{A^c}), P_{3,3}(\varphi_{A^c})\right)\right),
\end{aligned}$$

where the inequality follows by Poisson sampling, all the multiplicities are independent and hence $P(\varphi_A \cap \varphi_{A^c}) = P(\varphi_A)P(\varphi_{A^c})$, and also $\min(ab, cd) \geq \min(a, c)\min(b, d)$.

Since the product of the two summations is less than 2ϵ, at least one of them is less than $\sqrt{2\epsilon}$. If the second summation is less than $\sqrt{2\epsilon}$, then for any profile φ_{A^c}, if $P_{1,2}(\varphi_{A^c}) \geq \sqrt{2\epsilon}$ then for all R, $P_{3,3}(\varphi_{A^c}) < \sqrt{2\epsilon}$ and vice versa. This proves Condition 1.

If the first summation in the product is less than $\sqrt{2\epsilon}$, we bound it by the probability of types. Let type T_A denote the ordered set of joint multiplicities of elements in A. Clearly, $P(\varphi_A) = \sum_{T_A|\{T_A\}=\varphi_A} P(T_A)$. Since the $\min(\sum_i a_i, \sum_i b_i) \geq \sum_i \min(a_i, b_i)$,

$$\begin{aligned}\sum_{\varphi_A} \min\left(P_{1,2}(\varphi_A), P_{3,3}(\varphi_A)\right) &\geq \sum_{T_A} \min\left(P_{1,2}(T_A), P_{3,3}(T_A)\right)\\ &\overset{(a)}{\geq} \frac{1}{2E_{P_{1,2}}\left(\frac{P_{1,2}(T_A)}{P_{3,3}(T_A)}\right)}\\ &\overset{(b)}{\geq} \frac{1}{2}\exp\left(-\sum_{i\in A}\frac{n(p_i-q_i)^2}{p_i+q_i}\right),\end{aligned}$$

where (a) follows from Lemma 13. (b) follows from moment generating function of Poisson distributions and setting R to be $\frac{P+Q}{2}$. Condition 2 therefore holds. ■

The next lemma bounds the probability that the probability of a profile is less than ϵ.

Lemma 6 *Let X_1, X_2 be sequences of $\mathrm{Poi}(\mathrm{n}_0)$ elements generated i.i.d. according to $P_{1,2}$ and let A be the set of all symbols such that $n_0 p_i \leq 2\log n_0$, then*

$$\Pr(P_{1,2}(\varphi_A) \leq \epsilon) \leq \epsilon e^{120\log^3 n_0} + \frac{1}{100}.$$

Proof See Appendix B.3. ■

Next we show that $N_d(\varphi_A)$ and $N_s(\varphi_A)$ can be used to derive fairly good estimates of sub-profile probabilities. The proof is given in Appendix B.4.

Lemma 7 *For any profile φ_A and distributions $P_{1,2} = (P, Q)$, $P_{3,3} = (R, R)$*

$$P_{3,3}(\varphi_A), \leq \frac{N_s(\varphi_A)}{N_d(\varphi_A)} \leq \frac{P_{4,4}(\varphi_A)}{P_{1,2}(\varphi_A)},$$

where $P_{4,4} = (\frac{P+Q}{2}, \frac{P+Q}{2})$. ■

We can now prove the competitive optimality of the closeness test T of Algorithms 1 and the classifier S in Algorithm 2.

Theorem 8 *For every (P, Q),*

$$N^T_{P,Q} \leq \mathcal{O}({N^t_{P,Q}}^{1.5}\log N^t_{P,Q}) \qquad \textit{and} \qquad N^S_{P,Q} \leq \mathcal{O}({N^s_{P,Q}}^{1.5}\log N^s_{P,Q}).$$

Proof We first prove that if the sequences are generated from the same distribution then T returns `same` with probability ≥ 0.9, note that this part holds for every n. Similarly, we prove that if $N^t_{P,Q} \leq n$, then T returns `diff` with probability ≥ 0.9.

By Fact 12, with probability at least $1 - e^{-c_1 n^{3/2}/8}$, $n_1 + n_2 + n_3 \leq 4c_1 n^{3/2} \log n$. If $c_1 \geq 40$, then this probability is less than $1/60$. By the Poisson-tail bounds, $\Pr(\exists i \mid i \in A,\ p_i + q_i \leq 1/(2n\log^2 n)) \leq 1/120$. Similarly $\Pr(\exists i \mid i \in A^c,\ p_i + q_i \geq 2/(n\log^2 n)) \leq 1/120$. Throughout the proof, let μ_i and μ'_i denote the multiplicities of symbol i in sequences $\overline{x}_2$ and $\overline{y}_2$. Let φ_{A^c} denote the joint sub-profile of sequences $\overline{x}_3$ and $\overline{y}_3$.

We first prove that if the sequences are generated by the same distribution then T returns `same` with probability ≥ 0.9. Since both the sequences are generated by the same distribution, $\frac{(p_i - q_i)^2}{p_i + q_i} = 0$, for all $i \in A$. If all the probabilities in A are bigger than $1/(2n\log^2 n)$, then A contains at most $2n\log^2 n$ elements. By Chebyshev bound and Lemma 14 if $c_1 \geq 200$, then

$$\Pr\left(\sum_{i\in S} \frac{(\mu_i - \mu'_i)^2 - \mu_i - \mu'_i}{\mu_i + \mu'_i - 1} \geq \frac{c_1}{24}\sqrt{n}\log n\right) \leq \frac{1}{30},$$

Probability that **C2** holds is

$$\Pr\left(\frac{N_d(\varphi_{A^c})}{N_s(\varphi_{A^c})} \geq 34e^{120\log^3 n_0}\right) \overset{(a)}{\leq} \Pr\left(P_{3,3}(\varphi_{A^c}) \leq \frac{1}{34e^{120\log^3 n_0}}\right) \overset{(b)}{\leq} \frac{1}{30}.$$

(a) follows from Lemma 7 and (b) follows from Lemma 6. By the union bound, any of the conditions is satisfied is at most $1/60 + 1/60 + 1/30 + 1/30 = 1/10$.

We now prove that if the sequences are from (P, Q) such that $N^t_{P,Q} \leq n$, then T returns `diff` with probability ≥ 0.9. If $N^t_{P,Q} \leq n$, by Lemma 4, $N^t_{P,Q}(\epsilon' \overset{\text{def}}{=} \frac{1}{2}0.6^{c_2\log^3 n}) \leq c_2 n\log^3 n$. Hence, by Lemma 2, $\mathcal{E}^t_{P,Q}(\mathrm{Poi}(2n_0)) \leq \epsilon' + 2\mathrm{e}^{-\mathrm{c_2 n\log^3 n}/4} \leq 2\epsilon'$. Therefore, by Lemma 5 for any profile φ_A, if $P_{1,2}(\varphi_{A^c}) \geq \sqrt{2\epsilon'}$ then $P_{3,3}(\varphi_{A^c}) < \sqrt{2\epsilon'}$ and vice versa or for $c_2 \geq 14$

$$\sum_{i\in A} \frac{(p_i - q_i)^2}{p_i + q_i} \geq \frac{1}{4c_2 n\log^3 n} \log\left(\frac{1}{8(0.6)^{c_2\log^3 n}}\right) \geq \frac{1}{12n}.$$

If $c_1 \geq 6000$, then by Chebyshev bounds and Lemma 14

$$\Pr\left(\sum_{i\in A} \frac{(\mu_i - \mu'_i)^2 - \mu_i - \mu'_i}{\mu_i + \mu'_i - 1} \leq \frac{c_1}{24}\sqrt{n}\log n\right) \leq \frac{2}{30}.$$

If Condition 2 is satisfied in Lemma 5, then

$$\begin{aligned}
\Pr\left(\frac{N_d(\varphi_{A^c})}{N_s(\varphi_{A^c})} \leq 34e^{120\log^3 n_0}\right) &\overset{(a)}{\leq} \Pr\left(\frac{P_{1,2}(\varphi_{A^c})}{P_{4,4}(\varphi_{A^c})} \leq 34e^{120\log^3 n_0}\right) \\
&\overset{(b)}{\leq} \Pr(P_{1,2}(\varphi_{A^c}) \leq \sqrt{2\epsilon'}) + \Pr\left(P_{1,2}(\varphi_{A^c}) \leq \sqrt{68\epsilon'}e^{120\log^3 n_0}\right) \\
&\leq 2\sqrt{68\epsilon'}e^{240\log^3 n_0} + \frac{1}{100} \overset{(c)}{\leq} \frac{2}{30}.
\end{aligned}$$

(a) follows from Lemma 7. (b) follows from the fact that if $P_{4,4}(\varphi_A) \geq \epsilon'$, then $P_{1,2}(\varphi_A) \leq \epsilon'$. (c) holds when $c_2 \geq 300000$. By the union bound, probability the none of the conditions are satisfied at most $1/30 + 2/30 = 1/10$. By settting, $c_1 \geq c_2$, we obtain $c_1 n^{3/2} \log n \geq c_2 n \log n$.

Next we prove that if $n \geq N^s_{P,Q}$, Algorithm 2 classifies P and Q with error $\leq 1/10$. By Lemma 4, $N^s_{P,Q}(1/100) \leq 8N^s_{P,Q} \leq 8n$. Therefore, by Lemma 3, $N^t_{P,Q} \leq 40n$. Hence, if $\overline{z}$ and $\overline{x}$ are generated by same distribution and have length $c_3 n^{1.5} \log n \geq 900 c_1 n^{1.5} \log n$, then algorithm 1 returns `same` with probability $\geq 9/10$. Similarly if $\overline{z}$ and $\overline{y}$ are from same distribution then the algorithm returns `diff` with probability $\geq 9/10$. ■

We relate the sample complexity to achieve ϵ error to $N^t_{P,Q}$ is given by the following corollary.

Corollary 9 *For all (P, Q) and $\epsilon < 1/10$,*

$$N^T_{P,Q}(\epsilon) \leq \mathcal{O}\left({N^t_{P,Q}}^{1.5} \log N^t_{P,Q} \log \frac{1}{\epsilon}\right) \qquad \text{and} \qquad N^S_{P,Q}(\epsilon) \leq \mathcal{O}\left({N^s_{P,Q}}^{1.5} \log N^s_{P,Q} \log \frac{1}{\epsilon}\right).$$

Proof Follows from Lemma 4 and Theorem 8. ■

The algorithms can be slightly modified to improve the competitiveness for any given ϵ to $\mathcal{O}({N^t_{P,Q}}^{1.5}(\epsilon) \log N^t_{P,Q}(\epsilon))$ and $\mathcal{O}({N^s_{P,Q}}^{1.5}(\epsilon) \log^2 N^s_{P,Q}(\epsilon))$ respectively. Details are omitted for brevity.

6. Lower bounds on sample complexity

Using arguments similar to LeCam (1986); Paninski (2008), we prove the lower bounds by constructing a set of distributions that cannot all be simultaneously distinguished by a single algorithm. Let Q be the distribution over $i = 1, 2, \ldots, \frac{n^{1/3}}{\log n}$ where $q_i = \frac{3i^2 \log^3 n}{cn}$, and $c = \left(1 + \frac{\log n}{n^{1/3}}\right)\left(1 + \frac{\log n}{2n^{1/3}}\right)$ is a normalization factor. Let $t = \frac{n^{1/6}}{500 \log n}$, and define $\mathcal{P}$ to be the collection of $2^{n^{1/3}/2 \log n}$ distributions where in every $P \in \mathcal{P}$, for all odd i, $p_i = q_i \pm \frac{i \log n}{tn}$ and $p_{i+1} = q_{i+1} \mp \frac{i \log n}{tn}$, namely, $p_i + p_{i+1} = q_i + q_{i+1}$. The next lemma, proved in Appendix B.5, states that each distribution in $\mathcal{P}$ can be easily distinguished from Q.

Lemma 10 *For all $P \in \mathcal{P}$, $N^t_{P,Q} \leq n$ and $N^s_{P,Q} \leq n$.* ■

Theorem 11 *For every closeness test T and classifier S, there are distribution pairs (P, Q) (that may differ for the two cases) such that*

$$N^T_{P,Q} \geq \Omega\left(\frac{{N^t_{P,Q}}^{7/6}}{\log N^t_{P,Q}}\right) \qquad \text{and} \qquad N^S_{P,Q} \geq \Omega\left(\frac{{N^s_{P,Q}}^{7/6}}{\log N^s_{P,Q}}\right).$$

Proof We prove that no classifier S can classify Q from all $\mathcal{P}$ with $\Omega(n^{7/6}/\log n)$ samples. Since extra information decreases the error probability, we aid the symmetric classifier with about the label-probability mapping. Any such classifier divides $\mathcal{A}^n$ into two sets $\mathcal{A}_1$ and $\mathcal{A}_1^c$. If $\overline{z} \in \mathcal{A}_1$, then the classifier associates $\overline{z}$ with P, else Q. If Q can be distinguished from the set $\mathcal{P}$ with error P_e, then

$$P_e \geq \frac{1}{2} \min_{\mathcal{A}_1} \max_{P \in \mathcal{P}} Q(\mathcal{A}_1) + P(\mathcal{A}_1^c).$$

Similarly, for the closeness test T we provide additional information about labellings and further restrict that the first sequence is from P and the second sequence is either from P or Q, the error probability of such a closeness test is also lower bounded by the above equation. Hence, the error probability is lower bounded by,

$$P_e \geq \frac{1}{2}\min_{\mathcal{A}_1}\max_{P\in\mathcal{P}} Q(\mathcal{A}_1) + P(\mathcal{A}_1^c) \overset{(a)}{=} \frac{1}{2}\left(\min_{\mathcal{A}_1}\max_{\mu} Q(\mathcal{A}_1) + \sum_{P\in\mathcal{P}}\mu_P P(\mathcal{A}_1^c)\right)$$
$$\overset{(b)}{\geq} \frac{1}{2}\left(\max_{\mu}\min_{\mathcal{A}_1} Q(\mathcal{A}_1) + \sum_{P\in\mathcal{P}}\mu_P P(\mathcal{A}_1^c)\right) \overset{(c)}{\geq} \frac{1}{2}\max_{\mu}\sum_{\overline{z}}\min\left(Q(\overline{z}), \sum_{P\in\mathcal{P}}\mu_P P(\overline{z})\right),$$

where μ is any measure over $\mathcal{P}$. (a) follows from the fact that maximum of a linear optimization problem occurs at the boundary. Min-max is bigger than max-min, hence (b). (c) follows from the error probability of the optimal hypothesis test with known prior.

Since the inequality is true for any measure we choose μ to be uniformly over the set $\mathcal{P}$. Applying Lemma 13 to $R(\overline{Z})$ and $\sum_P \mu_P P(\overline{Z})$

$$P_e \geq \frac{1}{4E_{\mu_p P(\overline{Z})}\frac{\mu_p P(\overline{Z})}{Q(\overline{Z})}} \overset{(a)}{\geq} \frac{1}{4}\exp\left(-(n')^2\sum_i \delta_i^4\left(\frac{1}{q_{2i-1}^2}+\frac{1}{q_{2i}^2}\right)\right), \tag{3}$$

where $\delta_i = \frac{i\log n}{tn}$. Proof of (a) is given in Appendix B.6. Substituting the values,

$$P_e \geq \frac{1}{4}\exp\left(-16(n')^2\frac{n^{1/3}}{\log(n)}\frac{c^2}{9n^2t^4\log^2 n}\right) \geq \frac{1}{4}\exp\left(-16(n')^2\log(n)\frac{500^4c^2}{9n^{7/3}}\right).$$

For lower bound to be lesser than $1/10$, $n' = \Omega(n^{7/6}/\log n)$ samples are necessary. ■

References

J. Acharya, H. Das, H. Mohimani, A. Orlitsky, and Shengjun Pan. Exact calculation of pattern probabilities. In *Proceedings of IEEE Symposium on Information Theory*, pages 1498 –1502, 2010a.

J. Acharya, H. Das, A. Orlitsky, S. Pan, and N. P. Santhanam. Classification using pattern probability estimators. In *ISIT*, pages 1493–1497, 2010b.

J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. *Journal of Machine Learning Research - Proceedings Track*, 19:47–68, 2011.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *Annual IEEE Symposium on Foundations of Computer Science*, page 259, 2000.

T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. *Annual IEEE Symposium on Foundations of Computer Science*, page 442, 2001.

Tugkan Batu. *Testing properties of distributions*. PhD thesis, Cornell University, 2001.

G. Biau and L. Gyorfi. On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965 – 3973, nov. 2005.

A. Dhulipala and A. Orlitsky. Universal compression of markov and related sources over arbitrary alphabets. *IEEE Transactions on Information Theory*, 52:4182–4190, 2006.

S. Guha, A. McGregor, and S. Venkatasubramanian. Sublinear estimation of entropy and information distances. *ACM Trans. Algorithms*, 5:35:1–35:16, November 2009.

M. Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Transactions on Information Theory*, 35:401–408, 1989.

B. Kelly, T. Tularak, A. B. Wagner, and P. Viswanath. Universal hypothesis testing in the learning-limited regime. In *Proceedings of IEEE Symposium on Information Theory*, pages 1478–1482, 2010.

L. LeCam. *Asymptotic methods in statistical decision theory*. Springer series in statistics. Springer, New York, 1986.

M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang. On modeling profiles instead of values. In *UAI*, pages 426–435, 2004a.

A. Orlitsky, N.P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50:1469–1481, 2004b.

A. Orlitsky, S. Pan, Sajama, N.P. Santhanam, and K. Viswanathãn. Pattern maximum likelihood: computaiton and experiments. *In preparation*, 2012.

L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

S. Raskhodnikova. *Property Testing: Theory and Applications*. PhD thesis, Massachusetts Institute of Technology, 2003.

B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

G. Valiant and P. Valiant. The power of linear estimators. In *Annual IEEE Symposium on Foundations of Computer Science*, pages 403 –412, oct. 2011.

P. Valiant. Testing symmetric properties of distributions. In *ACM Symposium on Theory of Computing*, pages 383–392, New York, NY, USA, 2008. ACM.

Junan Zhang. *Universal Compression and Probability Estimation with Unknown Alphabets*. PhD thesis, UCSD, 2005.

J. Ziv. On classification with empirically observed statistics and universal data compression. *IEEE Transactions on Information Theory*, 34:278–286, 1988.

Appendix A. General probability bounds

We use the following Poisson-tail bound.

Fact 12 *If $X \sim \mathrm{Poi}(\lambda)$, then for $x \geq \lambda$,*

$$\Pr(X \geq x) \leq \exp\left(-\frac{(x-\lambda)^2}{2x}\right),$$

and for $x \leq \lambda$,

$$\Pr(X \leq x) \leq \exp\left(-\frac{(x-\lambda)^2}{2\lambda}\right). \qquad \blacksquare$$

Lemma 13 *For every (P, Q) over any alphabet $\mathcal{A}$,*

$$\sum_{z\in\mathcal{A}} \min(P(z), Q(z)) \geq \frac{1}{2E_P\left(\frac{P(z)}{Q(z)}\right)}.$$

Proof

$$\sum_z \min\left(P(z), Q(z)\right) \overset{(a)}{\geq} \sum_z \frac{P(z)Q(z)}{P(z)+Q(z)} = E_P\left(\frac{Q(z)}{P(z)+Q(z)}\right)$$
$$\overset{(b)}{\geq} \frac{1}{E_P\left(\frac{P(z)+Q(z)}{Q(z)}\right)} = \frac{1}{E_P\left(\frac{P(z)}{Q(z)}\right)+1} \overset{(c)}{\geq} \frac{1}{2E_P\left(\frac{P(z)}{Q(z)}\right)},$$

where (a) follows from $\min(a,b) \geq \frac{ab}{a+b}$, (b) follows from Jensen's inequality and the convexity of $\frac{1}{x}$, and (c) since by the Cauchy-Schwarz inequality, $E_P\left(\frac{P(z)}{Q(z)}\right) \geq 1$. $\blacksquare$

Lemma 14 *For two Poisson random variables, μ and μ' with means λ and λ' respectively,*

$$E\left(\frac{(\mu-\mu')^2-\mu-\mu'}{\mu+\mu'-1}\right) = \frac{(\lambda-\lambda')^2}{\lambda+\lambda'}\left(1-e^{-\lambda-\lambda'}\right),$$
$$\mathrm{Var}\left(\frac{(\mu-\mu')^2-\mu-\mu'}{\mu+\mu'-1}\right) \leq 4\frac{(\lambda-\lambda')^2}{\lambda+\lambda'} + 4 + \mathcal{O}((\lambda+\lambda')^5 e^{-\lambda-\lambda'}).$$

Proof

$$E\left(\frac{(\mu-\mu')^2-\mu-\mu'}{\mu+\mu'-1}\right) = E\left(\frac{\mu(\mu-1)+\mu'(\mu'-1)-2\mu\mu'}{\mu+\mu'-1}\right).$$

Analyzing the first term in the expression,

$$\begin{aligned}
E\left(\frac{\mu(\mu-1)}{\mu+\mu'-1}\right) &= e^{-\lambda-\lambda'}\lambda^2 \sum_{\mu=2,\mu'=0}^{\infty} \frac{\lambda^{\mu-2}\lambda'^{\mu'}}{(\mu-2)!\mu'!(\mu+\mu'-1)} \\
&= e^{-\lambda-\lambda'}\lambda^2 \sum_{\mu''=0,\mu'=0}^{\infty} \frac{\lambda^{\mu}\lambda'^{\mu''}}{(\mu'')!\mu'!(\mu+\mu'+1)} \\
&= \lambda^2 \sum_{\mu'''=0}^{\infty} e^{-\lambda-\lambda'} \frac{(\lambda+\lambda')^{\mu'''}}{(\mu''')!(\mu'''+1)} \\
&= \frac{\lambda^2}{\lambda+\lambda'} \sum_{\mu'''=0}^{\infty} e^{-\lambda-\lambda'} \frac{(\lambda+\lambda')^{\mu'''+1}}{(\mu'''+1)!} \\
&= \frac{\lambda^2}{\lambda+\lambda'}(1-e^{-\lambda-\lambda'}).
\end{aligned}$$

Similarly $E\left(\frac{\mu'(\mu'-1)}{\mu+\mu'-1}\right) = \frac{\lambda'^2}{\lambda+\lambda'}(1-e^{-\lambda-\lambda'})$ and $E\left(\frac{\mu\mu'}{\mu+\mu'-1}\right) = \frac{\lambda\lambda'}{\lambda+\lambda'}(1-e^{-\lambda-\lambda'})$, completing the expectation argument. The variance calculation can be done similarly by separating terms as in (a) and proving the upper bound,

$$\begin{aligned}
E\left(\frac{(\mu-\mu')^2-\mu-\mu'}{\mu+\mu'-1}\right)^2 &\overset{(a)}{=} (\lambda-\lambda')^4 E\left(\frac{1}{(\mu+\mu'+3)^2}\right) \\
&\quad + 4(\lambda-\lambda')^2(\lambda+\lambda')E\left(\frac{1}{(\mu+\mu'+2)^2}\right) + 2(\lambda+\lambda')^2 E\left(\frac{1}{(\mu+\mu'+1)^2}\right) \\
&\leq \frac{(\lambda-\lambda')^4}{(\lambda+\lambda')^2} + 4\frac{(\lambda-\lambda')^2}{\lambda+\lambda'} + 4 + \mathcal{O}((\lambda+\lambda')^5 e^{-\lambda-\lambda'}). \qquad \blacksquare
\end{aligned}$$

Appendix B. Proofs of lemmas in Sections 5 and 6

B.1. Proof of Lemma 3

To prove the upper bound, take an optimal closeness test that decides whether two sequences of $N^t_{P,Q}(\epsilon)$ samples are generated according to P and Q or by the same distribution, with error probability $\leq \epsilon$. Given classification sequences $\overline{X}$, $\overline{Y}$, and $\overline{Z}$, apply the test to $\overline{X}$ and $\overline{Z}$. If the test finds them to be generated according to the same distribution, classify $\overline{Z}$ to $\overline{X}$, and otherwise to $\overline{Y}$. The error probability is clearly $\leq \epsilon$.

To prove the lower bound, take an optimal classifier which xclassifies with error probability $\leq \epsilon$. Divide the $N^s_{P,Q}(\epsilon)\log(1/\epsilon)$ samples from both sequences into $\log(1/\epsilon)$ independent $N^s_{P,Q}(\epsilon)$ length sequences: $\overline{X}_1, \overline{X}_2, \ldots \overline{X}_{\log(1/\epsilon)}$ and $\overline{Y}_1, \overline{Y}_2, \ldots \overline{Y}_{\log(1/\epsilon)}$. Classify, $\overline{X}_{2i}$, $\overline{Y}_{2i}$ using $\overline{X}_{2i-1}, \overline{Y}_{2i-1}, \forall\, i$. If all the $\log(1/\epsilon)$ sequences are classified correctly, then declare `diff`, else `same`.

If the sequences are generated by the same distribution, then each of $\overline{X}_{2i}$ and $\overline{Y}_{2i}$ are classified as $\overline{X}_{2i-1}$ or $\overline{Y}_{2i-1}$ with probability 0.5. The error probability is the probability that all of them are

classified correctly, which is $2^{-\log(1/\epsilon)} = \epsilon$. If the sequences are from P and Q, then probability that any one of $\overline{X}_{2i}$ or $\overline{Y}_{2i}$ incorrectly classified is ϵ. Therefore, by the union bound the probability that any one of them is classified incorrectly is at most $\log(1/\epsilon)\epsilon$. The error probability of this test is $\epsilon \log(1/\epsilon)$, proving the lower bound. ■

B.2. Proof of Lemma 4

We prove the result for closeness testing. Similar result holds for classification and is omitted for brevity. Divide the two sequences of length nm into m sequences of length n. Run the closeness test for each one of them. If majority of m sequences declare `same`, declare `same`, else `diff`. Since each of m sub-tests fail with error probability $\mathcal{E}^t_{P,Q}(n)$, the error probability that majority test fails is at most

$$P_e \le \sum_{i=\lfloor \frac{m}{2} \rfloor + 1}^{m} \binom{m}{i} {\mathcal{E}^t_{P,Q}}^i (1 - \mathcal{E}^t_{P,Q})^{m-i} \le \left(\sum_{i=\lfloor \frac{m}{2} \rfloor + 1}^{m} \binom{m}{i} \right) \max_{i \ge \lfloor \frac{m}{2} \rfloor + 1} {\mathcal{E}^t_{P,Q}}^i (1 - \mathcal{E}^t_{P,Q})^{m-i}$$
$$\le \frac{1}{2} 2^m {\mathcal{E}^t_{P,Q}}^{m/2} (1 - \mathcal{E}^t_{P,Q})^{m/2} = \frac{1}{2} \left(2 {\mathcal{E}^t_{P,Q}}^{1/2} (1 - \mathcal{E}^t_{P,Q})^{1/2} \right)^m . \quad ■$$

B.3. Proof of Lemma 6

Since the distribution is sampled $\mathrm{Poi}(\mathrm{n}_0)$ times, the multiplicities are independent, and the Poisson-tail bounds in Mitzenmacher and Upfal (2005) yields,

$$\Pr\left(\mu_i \ge 7.7 \log n_0\right) \le \frac{1}{100 n_0}.$$

Since $\Pr(\max(X, Y) \ge \epsilon) \le \Pr(X + Y \ge \epsilon)$, grouping terms increases the probability. Group all symbols so that the sum of the λ's in each group is between $\log n_0$ and $2 \log n_0$. There are at most n_0 such groups, hence by the union bound,

$$\Pr\left(\bigcup_{i \in A} \mu_i \ge 7.7 \log n_0 \right) \le \frac{1}{100}.$$

Since, $\sum_i n p_i \le n_0$, the probability that $\sum_i \mu_i \ge 2 n_0$ is exponentially small. The number of profiles such that all multiplicities are less than $15 \log n_0$ and the sum of multiplicities is less than $2n$. There are at most $(7.7 \log n_0)^2$ prevalences and each one of them is at most $2 n_0$. Hence the total number of profiles with probability less than ϵ is at most $(2 n_0)^{(7.7 \log n_0)^2}$. Hence, the probability that $P_{1,2}(\varphi_{A}) \le \epsilon$ is upper bounded by,

$$\Pr(P_{1,2}(\varphi_{A}) \le \epsilon) \le \epsilon e^{120 \log^3 n_0} + \frac{1}{100}. \quad ■$$

B.4. Proof of Lemma 7

Recall that

$$P_{1,2}(\varphi_{A}) = \frac{\sum_\sigma \prod_{i \in A} e^{-np_i - nq_i} (np_i)^{\mu_{\sigma_i}} (nq_i)^{\mu'_{\sigma_i}}}{N_d(\varphi_{A})},$$

where the summation is over all symbol permutations. For a joint profile φ, let φ' be the *concatenated profile* whose multiplicities are the sum of the multiplicities of each element in φ. For example, if $\varphi = \{(0,1),(1,2),(2,1)\}$, then $\varphi' = \{1,3,3\}$. Again identifying φ'_{A} with the set of all concatenated sequences X^nY^n whose sub-profile corresponding to A is φ'_{A}, then

$$P_{P,P}(\varphi'_{A}) = \frac{\sum_{\sigma}\prod_{i\in A} e^{-2np_i}(np_i)^{\mu_{\sigma_i}}}{N_s(\varphi_{A})}.$$

The lower bound then follows as for all R

$$\frac{N_s(\varphi_{A})}{N_d(_{A})} = \frac{P_{3,3}(\varphi_{A})}{R(\varphi'_{A})} \geq P_{3,3}(\varphi_{A}).$$

For the upper bound, since the sum of a $\mathrm{Poi}(\lambda)$ and an independent $\mathrm{Poi}(\lambda')$ is $\mathrm{Poi}(\lambda+\lambda')$, we obtain $P_{P,Q}(\varphi'_{A}) = P_{\frac{P+Q}{2},\frac{P+Q}{2}}(\varphi'_{A})$, and since $\varphi_{A} \in \varphi'_{A}$, $P_{1,2}(\varphi_{A}) \leq P_{1,2}(\varphi'_{A})$. Combining all the above terms

$$\frac{N_s(\varphi_{A})}{N_d(\varphi_{A})} = \frac{P_{4,4}(\varphi_{A})}{P_{\frac{P+Q}{2},\frac{P+Q}{2}}(\varphi'_{A})} = \frac{P_{4,4}(\varphi_{A})}{P_{P,Q}(\varphi'_{A})} \leq \frac{P_{4,4}(\varphi_{A})}{P_{1,2}(_{A})}.$$ ■

B.5. Proof of Lemma 10

If we prove that $N^s_{P,Q}(1/1000) \leq n/10$, then by Lemma 3

$$N^t_{P,Q}(\log(1000)/1000) \leq \log(1000) N^s_{P,Q}(1/1000) \leq n.$$

We construct a symmetric classifier with prior knowledge of P and Q to prove $N^s_{P,Q}(1/1000) \leq n/10$. The classifier selects first $\mathrm{Poi}(\mathrm{n}/20)$ samples from $\overline{z}$, denoted by $\overline{z}'$. It assigns symbol i to the symbol with i^{th} highest multiplicity. It assigns $\overline{z}$ to the distribution that assigns higher probability. Let H be the event that the probabilities are assigned to symbols incorrectly. The error probability of the classifier is upper bounded by

$$P_e \leq \Pr(H^c) + \frac{1}{2}\Pr(H)\left(\Pr_P(P(\overline{Z}') \leq Q(\overline{Z}') \mid H) + \Pr_Q(Q(\overline{Z}') \leq P(\overline{Z}') \mid H)\right).$$

By Poisson tail bounds 12 $\Pr(H) \leq \mathrm{poly}(1/\mathrm{n})$. Furthermore $\Pr_P(P(\overline{Z}') \leq Q(\overline{Z}') \mid H) = \Pr_P(P(\overline{Z}) \leq Q(\overline{Z}))$. Note that

$$\Pr_P(P(\overline{Z}) \leq Q(\overline{Z})) + \Pr_Q(P(\overline{Z}) \leq Q(\overline{Z})) = \sum_{\overline{z}} \min(P(\overline{z}), Q(\overline{z})).$$

Therefore

$$P_e - \mathrm{poly}(1/\mathrm{n}) \leq \frac{1}{2}\sum_{\overline{\mathrm{z}}} \min(\mathrm{P}(\overline{\mathrm{z}}), \mathrm{Q}(\overline{\mathrm{z}})) \leq \frac{\sum_{\overline{\mathrm{z}}}\sqrt{\mathrm{P}(\overline{\mathrm{z}})\mathrm{Q}(\overline{\mathrm{z}})}}{2} \overset{(\mathrm{a})}{=} \frac{1}{2}\sum_{\mathrm{i}} \exp\left(-\frac{\mathrm{n}}{20}(\sqrt{\mathrm{p_i}} - \sqrt{\mathrm{q_i}})^2\right)$$

$$\leq \frac{1}{2}\exp\left(-\frac{n}{20}\sum_i \frac{(p_i-q_i)^2}{4q_i}\right) \leq \frac{1}{2}\exp\left(-\frac{n^{1/3}c}{480\log^2(n)t^2}\right),$$

where (a) follows from the moment generating functions of Poisson distributions and by Fact 1. If $t = \frac{n^{1/6}}{500\log(n)}$, then all each of $P \in \mathcal{P}$, $N^s_{P,Q}(1/1000) \leq n/10$. ■

B.6. Proof of Equation (3)

Let $\delta_i = \frac{(2i-1)}{nt}$ and $\mathrm{Poi}(\lambda, \mathrm{i}) = \mathrm{e}^{-\lambda}\frac{\lambda^{\mathrm{i}}}{\mathrm{i!}}$ Since the selected mixture is uniform, it can be decomposed into product over individual symbols. Therefore, $E_{\mu_P P(\overline{Z})}\left(\frac{\mu_P P(\overline{Z})}{Q(\overline{Z})}\right)$ is given by

$$
\begin{aligned}
&= \prod_i \sum_{j,k} \left(\frac{\left(\sum_{t=\{-1,1\}} \mathrm{Poi}(\mathrm{n}(\mathrm{q}_{2\mathrm{i}-1} + \mathrm{t}\delta_{\mathrm{i}}), \mathrm{j})\mathrm{Poi}(\mathrm{n}(\mathrm{q}_{2\mathrm{i}} - \mathrm{t}\delta_{\mathrm{i}}), \mathrm{k})\right)^2}{4\mathrm{Poi}(\mathrm{nq}_{2\mathrm{i}-1}, \mathrm{j})\mathrm{Poi}(\mathrm{nq}_{2\mathrm{i}}, \mathrm{k})} \right) \\
&\overset{(a)}{=} \prod_i \frac{1}{2} \exp\left(\delta_i^2 \left(\frac{1}{nq_{2i-1}} + \frac{1}{nq_{2i}}\right)\right) + \exp\left(-\delta_i^2 \left(\frac{1}{nq_{2i-1}} + \frac{1}{nq_{2i}}\right)\right) \\
&\overset{(b)}{\leq} \prod_i \exp\left(\frac{1}{2} n^2 \delta_i^4 \left(\frac{1}{q_{2i-1}} + \frac{1}{q_{2i}}\right)^2\right) \\
&\overset{(c)}{\leq} \exp \sum_i \left(n^2 \delta_i^4 \left(\frac{1}{q_{2i-1}^2} + \frac{1}{q_{2i}^2}\right)\right),
\end{aligned}
$$

where (a) follows from by evaluating the expressions using moment generating functions of Poisson distributions, (b) follows from $\frac{1}{2}(e^t + e^{-t}) \leq e^{t^2/2}$, and (c) follows from the AM-GM Inequality.