Competitive Closeness Testing

Jayadev Acharya Hirakendu Das Ashkan Jafarpour Alon Orlitsky Shengjun Pan
University of California San Diego
La Jolla, CA 92093
{jacharya,hdas,ajafarpo,alon,s1pan}@ucsd.edu

Abstract

We test whether two sequences are generated by the same distribution or by two different ones. Unlike previous work, we make no assumptions on the distributions' support size. Additionally, we compare our performance to that of the best possible test. We describe an efficiently-computable algorithm based on *pattern* maximum likelihood that is near optimal whenever the best possible error probability is $\le \exp(-14n^{2/3})$ using length-n sequences.

1 Introduction

We consider the problem of testing whether two sequences are generated by the same distribution or by two different ones. There is an extensive amount of literature on this problem and several of its variants in the framework of hypothesis testing [6, 23, 7, 11, 12], which primarily considers asymptotic error performance when the sequence lengths tend to infinity.

For non-asymptotic lengths, significant progress has been made recently under distribution property testing [3, 4, 18, 21], which provide efficient algorithms for closeness testing and other problems like entropy estimation and support size estimation using a number of samples that are sublinear in the support size. Nonetheless, these algorithms and their error performance guarantees require a priori knowldedge of upper bounds on the support size. In this paper, we present closeness-testing algorithms that are competitively optimal when the best possible error probability is small. The algorithms do not require knowledge of the underlying support size. Our methods extend the technique of pattern maximum likelihood (PML) used in [16, 15] for estimating large alphabet distributions in the context of universal compression.

1.1 Problem definition

Let $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$ be an alphabet of size k and (p_1, p_2) be a pair of unknown distributions over $\mathcal{A}$. Two length-n sequences $\overline{X}_1, \overline{X}_2 \in \mathcal{A}^n$ are generated *i.i.d.* and independently of each other according to p_1 and p_2 respectively. The problem is to decide whether p_1 and p_2 are same or different given only $\overline{X}_1$ and $\overline{X}_2$ and no further information about (p_1, p_2). A closeness test Δ for sequences in $\mathcal{A}^n$ is a mapping $\Delta : \mathcal{A}^n \times \mathcal{A}^n \to \{\textit{same}, \textit{diff}\}$ that labels each sequence pair as *same* or *diff*, indicating whether the distributions that generated them are same or different. The error probability of Δ for any (p_1, p_2) is the probability that it labels a sequence pair generated by (p_1, p_2) incorrectly, *i.e.,*

$$P_{\mathrm{e}}^n(\Delta, p_1, p_2) \stackrel{\mathrm{def}}{=} \begin{cases} \Pr(\Delta(\overline{X}_1, \overline{X}_2) = \textit{diff}) & \text{if } p_1, p_2 \text{ are same,} \\ \Pr(\Delta(\overline{X}_1, \overline{X}_2) = \textit{same}) & \text{if } p_1, p_2 \text{ are different.} \end{cases}$$

The goal is to design a test Δ that uses few samples and yet has a low error probability, both when (p_1, p_2) are same, *i.e.,* $p_1 = p_2$ and when (p_1, p_2) are sufficiently different to be distinguishable by some test.

1.2 A closeness test based on empirical distributions

As noted in [3], the problem can be regarded as a composite hypothesis testing problem [17] where there are two classes of distribution pairs, $\mathcal{P}_{\textit{same}}$ that contains pairs of identical distributions (p, p), and $\mathcal{P}_{\textit{diff}}$ that contains pairs of significantly different distributions (p_1, p_2). For simple hypothesis testing problems (where there is only one distribution, or one distribution pair in our case, in each class), a likelihood ratio test (LRT) [6, 17] has lowest error probability. The empirical frequency distribution, which is also the maximum likelihood distribution of the sequence, is a good estimate of the underlying distribution when alphabet size k is small and sequence length n is large. Hence, one can plug in the empirical distributions instead of actual distributions into a LRT to obtain a test that has low error probability in this case. Such a ratio test that uses maximum likelihood distributions is commonly referred to as a generalized likelihood ratio test (GLRT) [17] and is often used for composite hypothesis testing.

Specifically, let $\hat{P}(\overline{x}) \overset{\text{def}}{=} \max_p p(\overline{x}) = \prod_{a \in \mathcal{A}} \left(\frac{\mu(a)}{n}\right)^{\mu(a)}$ be the maximum likelihood of a sequence $\overline{x} \in \mathcal{A}^n$ under all possible *i.i.d.* distributions, where $\mu(a)$ is the number of appearances of a symbol a in $\overline{x}$. Define Δ^{emp} as the test such that

$$\Delta^{\text{emp}}(\overline{x}_1, \overline{x}_2) \overset{\text{def}}{=} \begin{cases} \textit{diff} & \text{if } \frac{\hat{P}(\overline{x}_1)\hat{P}(\overline{x}_2)}{\hat{P}(\overline{x}_1\overline{x}_2)} > \binom{n+k-1}{n}^2 n, \\ \textit{same} & \text{otherwise,} \end{cases}$$

for all $(\overline{x}_1, \overline{x}_2) \in \mathcal{A}^n \times \mathcal{A}^n$. Note that for all $(\overline{x}_1, \overline{x}_2)$, $\hat{P}(\overline{x}_1)\hat{P}(\overline{x}_2)/\hat{P}(\overline{x}_1\overline{x}_2) \geq 1$ since

$$\hat{P}(\overline{x}_1)\hat{P}(\overline{x}_2) = \max_{p_1,p_2} p_1(\overline{x}_1)p_2(\overline{x}_2) \geq \max_{p_1=p_2} p_1(\overline{x}_1)p_2(\overline{x}_2) = \hat{P}(\overline{x}_1\overline{x}_2).$$

It can be shown that when $p_1 = p_2$, $\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)/\hat{P}(\overline{X}_1\overline{X}_2)$ is small and $\leq \binom{n+k-1}{n}^2 n$ with probability $\geq 1 - \frac{1}{n}$. And when $|p_1 - p_2| > \epsilon$ for some $\epsilon > 0$, then $\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)/\hat{P}(\overline{X}_1\overline{X}_2)$ is large and $\geq 2^{n\epsilon^2/6} > \binom{n+k-1}{n}^2 n$ with probability $1 - o(1)$ [10] when $k = o(n)$. Hence, when $k = o(n)$, *i.e.,* the alphabet size k is small and sublinear in n, then Δ^{emp} has low error probability, both in the case when $p_1 = p_2$ and in the case when the L_1 distance $|p_1 - p_2| > \epsilon$ for some constant $\epsilon > 0$.

However, when the alphabet size is large, empirical distribution may not be a good estimate of the underlying distribution and Δ^{emp} may not have low error probability, as shown in an example in [10] and in the following, simpler, example.

Example 1. Let $k = n^3$ for large n and let (p_1, p_2) be such that $p_1(a_1) = 1$, $p_2(a_1) = 1/2$ and for $i = 2, \ldots, k$, $p_1(a_i) = 0$, $p_2(a_i) = 1/(2(k-1))$. The two distributions are clearly very different and $|p_1 - p_2| = 1$. If $\overline{X}_1$ and $\overline{X}_2$ are length-n sequences generated *i.i.d.* according to p_1 and p_2 respectively, then $\overline{X}_1 = a_1^n$ and $\overline{X}_2 = a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1}$ are typical sequences. In particular, by the Birthday problem, no symbol in $\{a_2, a_3, \ldots, a_k\}$ appears more than once in $\overline{X}_2$ with high probability. In this case, we see that

$$\frac{\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)}{\hat{P}(\overline{X}_1\overline{X}_2)} = \frac{\hat{P}(a_1^n)\hat{P}(a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1})}{\hat{P}(a_1^{\frac{3n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1})} = \frac{1^n \times (\frac{1}{2})^{\frac{n}{2}} (\frac{1}{n})^{\frac{n}{2}}}{(\frac{3}{4})^{\frac{3n}{2}} (\frac{1}{2n})^{\frac{n}{2}}} = \left(\frac{4}{3}\right)^{\frac{3n}{2}} \approx 1.54^n.$$

When both $\overline{X}_1$ and $\overline{X}_2$ are generated according to p_2, $\overline{X}_1 = a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1}$ and $\overline{X}_2 = a_1^{\frac{n}{2}} a_{\frac{n}{2}+2} \cdots a_{n+1}$ are typical sequences and no symbol in $\{a_2, a_3, \ldots, a_k\}$ appears more than once in $\overline{X}_1\overline{X}_2$ with high probability. Then,

$$\frac{\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)}{\hat{P}(\overline{X}_1\overline{X}_2)} = \frac{\hat{P}(a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1})\hat{P}(a_1^{\frac{n}{2}} a_{\frac{n}{2}+2} \cdots a_{n+1})}{\hat{P}(a_1^n a_2 a_3 \cdots a_{n+1})} = \frac{(\frac{1}{2})^{\frac{n}{2}} (\frac{1}{n})^{\frac{n}{2}} \times (\frac{1}{2})^{\frac{n}{2}} (\frac{1}{n})^{\frac{n}{2}}}{(\frac{1}{2})^n (\frac{1}{2n})^n} = 2^n.$$

Hence, $\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)/\hat{P}(\overline{X}_1\overline{X}_2)$ is higher in the case when the distributions are same compared to when the distributions are different. Therefore, the test $\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)/\hat{P}(\overline{X}_1\overline{X}_2) \underset{\textit{same}}{\overset{\textit{diff}}{\gtrless}} t$ cannot have low error probability for both (p_1, p_2) and (p_2, p_2) for any choice of threshold t. Furthermore, we note that when $\overline{X}_1, \overline{X}_2$ are both generated according to p_2, $\overline{X}_1$, $\overline{X}_2$ and hence $\overline{X}_1\overline{X}_2$ have very different empirical distribution estimates in the computation of likelihood ratio. □

1.3 Related work on estimating large alphabet distributions

Batu et al [3, 4] developed a test that distinguishes the cases that two distributions are close or well separated in L_1 distance using sequences whose length is sublinear in size of the underlying alphabet. They show an algorithm that outputs ***pass*** (*i.e.,* ***same***) when $|p_1 - p_2| \leq \max(\frac{\epsilon}{32k^{1/3}}, \frac{\epsilon}{4k^{1/2}})$ and ***fail*** (*i.e.,* ***diff***) when $|p_1 - p_2| > \epsilon$ with error probability $\leq \delta$ in both cases, using sequences of length $n = \mathcal{O}(k^{2/3} \log k \cdot \epsilon^{-4} \cdot \log \frac{1}{\delta})$. Their algorithm estimates the L_1 distance contribution of only the high probability symbols using their relative frequencies as probability estimates, since the empirical distribution is a good probability estimate for such symbols. The contribution of low probability symbols is estimated using a test for L_2 distance that relies on the number of collisions (also known as coincidences or repetitions) in the sequences. They also show a corresponding lower bound by showing pairs of distributions (p_1, p_2) such that $|p_1 - p_2| > \epsilon$ and that no algorithm can distinguish it from the identical pair (p_1, p_1) using $n = o(k^{2/3} \cdot \epsilon^{-2/3})$ samples. Valiant [21] further showed that distinguishing distribution pairs with L_1 distance less than α from those with distance greater than β for $0 < \alpha < \beta < 2$ requires $n = k^{1-o(1)}$ samples and can be done using $n = \tilde{\mathcal{O}}(k)$ samples by [3] or by another test shown in [21]. Also see [4, 2, 5, 19, 21] and references therein for other problems of testing properties of distributions using a number of samples that is sublinear in the alphabet size. Although no assumptions are made on the structure of distributions, the tests in [3, 21] and their sample complexities still depend on the knowledge of an upper bound on the alphabet size k of the unknown underlying distributions. Moreover, as in Example 1, there are many distribution pairs that can be tested for closeness in much less than $\tilde{\mathcal{O}}(k^{2/3})$ samples.

The related problem of classification was considered by Kelly et al [10]. Here, one is given training sequences $\overline{X}_1$ and $\overline{X}_2 \in \mathcal{A}^n$ generated *i.i.d.* and independently according to unknown distributions p_1 and p_2 that are seperated in L_1 distance. A third sequence $\overline{Y} \in \mathcal{A}^n$ is generated *i.i.d.* and independently according to either p_1 or p_2 with equal probability and the problem is to decide whether $\overline{Y}$ is generated according p_1 or p_2. They show a test that has low error probability when (p_1, p_2) belong to a restricted class of distributions such that the probabilities of all symbols are $\Theta(\frac{1}{k})$ and $k = \Theta(n^\alpha)$, for any fixed $\alpha \in [0, 2)$. Their test uses the L_2 distance between the empirical frequency distributions, of the sequences to determine which one of the pairs $(\overline{X}_1, \overline{Y})$ or $(\overline{X}_2, \overline{Y})$ are closer and classify accordingly.

The problem of estimating the probability multiset of large alphabet distributions was also studied in the context of universal compression of large alphabet sources in [16, 14, 15]. The main idea is to consider the *pattern* of a sequence, which conveys only the structure of the sequence and the order in which symbols appear in the sequence, and not the identities of the actual symbols. The pattern contains all the information that is needed to test symmetric properties like entropy that depend only on the probability multiset and not on the way in which the probabilities are associated with the symbols of the alphabet. In [16], several estimators based on the maximum likelihood of patterns were shown that estimate the pattern probabilities (that are usually exponentially small in n) to within a factor that is subexponential in the sequence length n, regardless of the alphabet size and the structure of the underlying distribution. Preliminary results on application of such estimators to the problem of classification were shown in [20]. Partial results on classifiers based on maximum likelihood estimation of the *joint pattern* of two or more sequences were shown in [1]. In this paper, we show closeness tests based on maximum likelihood of joint patterns that perform almost as good as any test can, without making any assumptions on the underlying distributions. These tests can be used as good classifiers as well.

1.4 Closeness tests based on pattern maximum likelihood

The pattern of a sequence was introduced in [16]. Let $\overline{x} = x_1 x_2 \cdots x_n = x_1^n \in \mathcal{A}^n$ be a sequence of length n and $\mathcal{A}(\overline{x})$ denote the set of symbols that appear in $\overline{x}$. The index $\imath_{\overline{x}}(a)$ of a symbol $a \in \mathcal{A}(\overline{x})$ is

$$\imath_{\overline{x}}(a) \stackrel{\text{def}}{=} \min\{|\mathcal{A}(x_1^i)| : 1 \le i \le n \text{ and } x_i = a\},$$

i.e., one more than the number of distinct symbols that have appeared before the first appearance of a in $\overline{x}$. The *pattern* of $\overline{x}$ is the sequence

$$\Psi(\overline{x}) \stackrel{\text{def}}{=} \imath_{\overline{x}}(x_1)\imath_{\overline{x}}(x_2)\cdots\imath_{\overline{x}}(x_n)$$

obtained by replacing the symbols in $\overline{x}$ by their respective indices. For example, if $\overline{x} = \texttt{abracadabra}$, then $\imath_{\overline{x}}(\texttt{a}) = 1$, $\imath_{\overline{x}}(\texttt{b}) = 2$, $\imath_{\overline{x}}(\texttt{r}) = 3$, $\imath_{\overline{x}}(\texttt{c}) = 4$ and $\imath_{\overline{x}}(\texttt{d}) = 5$. Hence, $\Psi(\texttt{abracadabra}) = 12314151231$. The set of all possible patterns of different length-n sequences is represented by Ψ^n. For example, $\Psi^1 = \{1\}$, $\Psi^2 = \{11, 12\}$ and $\Psi^3 = \{111, 112, 121, 122, 123\}$.

We extend the definition of patterns to two or more sequences. The *joint pattern* of a pair of sequences $(\overline{x}_1, \overline{x}_2) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$ is $\Psi(\overline{x}_1, \overline{x}_2) \stackrel{\text{def}}{=} (\overline{\psi}_1, \overline{\psi}_2)$, where $\overline{\psi}_1 = \Psi(\overline{x}_1)$ and $\overline{\psi}_1\overline{\psi}_2 = \Psi(\overline{x}_1\overline{x}_2)$. For example, the joint pattern of $\texttt{bab}$ and $\texttt{abca}$ is $\Psi(\texttt{bab}, \texttt{abca}) = (121, 2132)$, since the pattern of first sequence is $\Psi(\texttt{bab}) = 121$ and that of the two sequences catenated is $\Psi(\texttt{bababca}) = 1212132$. Thus, the joint pattern conveys the patterns of the individual sequences and the association between the symbols of the sequences. The joint pattern of a list of three or more sequences is defined similarly.

We use Ψ^{n_1,n_2} to denote the set of all possible joint patterns of pairs of sequences of length (n_1, n_2). For example, $\Psi^{2,1} = \{(11,1),(11,2),(12,1),(12,2),(12,3)\}$.

The probability of a single pattern $\overline{\psi} \in \Psi^n$ under a distribution p is the probability that a length-n sequence $\overline{X}$ generated *i.i.d.* according to p has pattern $\overline{\psi}$, *i.e.,*

$$p(\overline{\psi}) \overset{\text{def}}{=} p\Big(\Psi(\overline{X}) = \overline{\psi}\Big) = \sum_{\overline{x}:\Psi(\overline{x})=\overline{\psi}} p(\overline{x}).$$

Similarly, the probability of a joint pattern $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n_1,n_2}$ under a pair of distributions (p_1, p_2) is the probability that two sequences $\overline{X}_1$ and $\overline{X}_2$ of length n_1 and n_2 generated *i.i.d.* according to p_1 and p_2 respectively have joint pattern $(\overline{\psi}_1, \overline{\psi}_2)$ and is denoted by

$$p_{1,2}(\overline{\psi}_1, \overline{\psi}_2) = p_{1,2}\Big(\Psi(\overline{X}_1, \overline{X}_2) = (\overline{\psi}_1, \overline{\psi}_2)\Big) = \sum_{\substack{(\overline{x}_1,\overline{x}_2): \\ \Psi(\overline{x}_1,\overline{x}_2)=(\overline{\psi}_1,\overline{\psi}_2)}} p_1(\overline{x}_1)p_2(\overline{x}_2).$$

For example, if $\mathcal{A} = \{\mathtt{a},\mathtt{b},\mathtt{c},\mathtt{d}\}$ and $p = (p_{\mathtt{a}}, p_{\mathtt{b}}, p_{\mathtt{c}}, p_{\mathtt{d}})$, then probability of the pattern $\overline{\psi} = 1213$ is

$$p(1213) = p(\mathtt{abac}) + p(\mathtt{abad}) + p(\mathtt{acab}) + \cdots = p_{\mathtt{a}}^2 p_{\mathtt{b}} p_{\mathtt{c}} + p_{\mathtt{a}}^2 p_{\mathtt{b}} p_{\mathtt{d}} + p_{\mathtt{a}}^2 p_{\mathtt{c}} p_{\mathtt{b}} + \cdots .$$

Similarly, if $p_1 = (p_{\mathtt{a}}, p_{\mathtt{b}}, p_{\mathtt{c}}, p_{\mathtt{d}})$ and $p_2 = (p'_{\mathtt{a}}, p'_{\mathtt{b}}, p'_{\mathtt{c}}, p'_{\mathtt{d}})$, then probability of the pattern $(12, 13)$ is

$$p_{1,2}(12,13) = p_{1,2}(\mathtt{ab},\mathtt{ac}) + p_{1,2}(\mathtt{ab},\mathtt{ad}) + p_{1,2}(\mathtt{ba},\mathtt{bc}) + \cdots = p_{\mathtt{a}} p_{\mathtt{b}} p'_{\mathtt{a}} p'_{\mathtt{c}} + p_{\mathtt{a}} p_{\mathtt{b}} p'_{\mathtt{a}} p'_{\mathtt{d}} + \cdots .$$

Notice that if $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n_1,n_2}$, then $\overline{\psi}_1\overline{\psi}_2 \in \Psi^{n_1+n_2}$. Also, if $p_1 = p_2 = p$, then $p_{1,2}(\overline{\psi}_1, \overline{\psi}_2) = p_{1,1}(\overline{\psi}_1, \overline{\psi}_2) = p_1(\overline{\psi}_1\overline{\psi}_2)$.

The maximum likelihood of a pattern $\overline{\psi}$ under all *i.i.d.* distributions is $\hat{P}(\overline{\psi}) \overset{\text{def}}{=} \max\limits_{p} p(\overline{\psi})$. Similarly, the maximum likelihood of a joint pattern $(\overline{\psi}_1, \overline{\psi}_2)$ under all pairs of *i.i.d.* and independent distributions is denoted by $\hat{P}(\overline{\psi}) \overset{\text{def}}{=} \max\limits_{p_1,p_2} p_{1,2}(\overline{\psi})$.

Since joint patterns contain all the relevant information for closeness testing, consider a simple hypothesis testing problem where a sequence pair $(\overline{X}_1, \overline{X}_2) \in \mathcal{A}^n \times \mathcal{A}^n$ is generated according to either according to (p_1, p_2) or (p, p), but we are given only the joint pattern $\Psi(\overline{X}_1, \overline{X}_2)$ and not the actual sequences. In this case, the likelihood ratio test $p_{1,2}(\Psi(\overline{X}_1, \overline{X}_2)) \underset{\mathit{same}}{\overset{\mathit{diff}}{\gtrless}} p(\Psi(\overline{X}_1\overline{X}_2))$ is a test with minimum error probability. Hence, similar to Subsection 1.2, viewing closeness testing as a composite hypothesis testing problem with the joint pattern of the sequences given as the observations, we consider the test $\Delta^{\hat{P}(\Psi)} \overset{\text{def}}{=} \Delta^{\hat{P}(\Psi)}_{n,\delta}$ defined as

$$\Delta^{\hat{P}(\Psi)}_{n,\delta}(\overline{x}_1, \overline{x}_2) \overset{\text{def}}{=} \begin{cases} \mathit{diff} & \text{if } \frac{\hat{P}(\Psi(\overline{x}_1,\overline{x}_2))}{\hat{P}(\Psi(\overline{x}_1\overline{x}_2))} > \frac{1}{\sqrt{\delta}}, \\ \mathit{same} & \text{otherwise,} \end{cases}$$

for all $(\overline{x}_1, \overline{x}_2) \in \mathcal{A}^n \times \mathcal{A}^n$ and for some $\delta < \exp(-12n^{2/3})$. In other words, the test outputs ***diff*** if the maximum likelihood of the pattern of the two sequences under two different distributions is much higher than that under two identical distributions.

Our first main result, Theorem 7, is that the test $\Delta^{\hat{P}(\Psi)}$ has low error probability when the pair of distributions is identical and also when the pair of distributions is significantly different such that there exists a test that can distinguish it from any pair of identical distributions. We note that, without loss of generality, we limit ourselves to only *symmetric* tests in our analysis, whose output depends only on joint pattern of the sequences and not the specific symbols that have appeared, since the property of closeness depends only on the probability multiset and not the associated symbols. (Also see Appendix C that provides a discussion along the lines of [3, 4].) We say that a pair of distributions (p_1, p_2) is (n, δ)-different if there exists a symmetric test that can distinguish with error probability $< \delta$, pairs of length n sequences generated according to (p_1, p_2) from those generated by any pair of identical distributions (p, p). In other words, for all p, there exists a test Δ such that

$$P_{\text{e}}^n(\Delta, p_1, p_2) < \delta \quad\text{and}\quad P_{\text{e}}^n(\Delta, p, p) < \delta.$$

Revisiting Example 1, in the case when $(\overline{X}_1, \overline{X}_2) \sim (p_1, p_2)$, consider the typical sequence pair $(\overline{X}_1, \overline{X}_2) = (a_1^n, a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1})$. Then, $\hat{P}(\Psi(\overline{X}_1, \overline{X}_2)) = \hat{P}(1^n, 1^{\frac{n}{2}} 23 \cdots (\frac{n}{2}+1)) \geq 1 \cdot (\frac{1}{2})^{\frac{n}{2}} (\frac{1}{2})^{\frac{n}{2}} =$

$(\frac{1}{2})^n$, since the distributions (p'_1, p'_2) assign $\Psi(\overline{X}_1, \overline{X}_2)$ such a likelihood, where $p'_1(a_1) = 1$, $p'_2(a_1) = \frac{1}{2}$, and the remaining probability $\frac{1}{2}$ of p'_2 is spread over a continous alphabet or a large tail, similar to p_2. Also, from [13], $\hat{P}(\Psi(\overline{X}_1\overline{X}_2)) = \hat{P}(1^{\frac{3n}{2}}23\cdots(\frac{n}{2}+1)) = (\frac{3}{4})^{\frac{3n}{2}}(\frac{1}{4})^{\frac{n}{2}}$, assigned by the distribution p such that $p(a_1) = \frac{3}{4}$ and has the remaining probability $\frac{1}{4}$ spread over a continuous alphabet. Hence,

$$\frac{\hat{P}(\Psi(\overline{X}_1, \overline{X}_2))}{\hat{P}(\Psi(\overline{X}_1\overline{X}_2))} \geq \frac{(\frac{1}{2})^n}{(\frac{3}{4})^{\frac{3n}{2}}(\frac{1}{4})^{\frac{n}{2}}} = \left(\frac{4}{3}\right)^{\frac{3n}{2}} > 1.53^n,$$

and the test $\Delta^{\hat{P}(\Psi)}$ outputs ***diff*** for $\delta = \exp(-14n^{2/3})$. When $(\overline{X}_1, \overline{X}_2) \sim (p_2, p_2)$, for the typical sequence pair $(\overline{X}_1, \overline{X}_2) = (a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1}, a_1^{\frac{n}{2}} a_{\frac{n}{2}+2} \cdots a_{n+1})$, again by [13], we have $\hat{P}(\Psi(\overline{X}_1, \overline{X}_2)) \leq \hat{P}(\Psi(\overline{X}_1)\hat{P}(\Psi(\overline{X}_2)) = \hat{P}(1^{\frac{n}{2}}23\cdots(\frac{n}{2}+1))^2 = \left((\frac{1}{2})^{\frac{n}{2}}(\frac{1}{2})^{\frac{n}{2}}\right)^2 = (\frac{1}{2})^{2n}$, and $\hat{P}(\Psi(\overline{X}_1\overline{X}_2)) = \hat{P}(1^n 23\cdots(n+1)) = (\frac{1}{2})^n(\frac{1}{2})^n = (\frac{1}{2})^{2n}$. Hence, in this case

$$\frac{\hat{P}(\Psi(\overline{X}_1, \overline{X}_2))}{\hat{P}(\Psi(\overline{X}_1\overline{X}_2))} = 1,$$

and the output of $\Delta^{\hat{P}(\Psi)}$ is ***same***. We note that the maximum likelihood distributions of $\Psi(\overline{X}_1, \overline{X}_2)$ and of $\Psi(\overline{X}_1\overline{X}_2)$ are consistent, *i.e.,* same, unlike in the case of Δ^{emp}.

As evident from the previous example, the computation of pattern maximum likelihood (PML) is difficult in general and hence we show an efficient test based on pattern probability estimators that also has low error probability. Several such estimators were shown in [16], that can compute maximum likelihood of patterns to within a subexponential factor. In particular, we consider the following estimator in [16]. The *profile* of a pattern or a sequence conveys the number of symbols appearing a given number of times in it. For example, the profile of `abdb` is $\varphi(\texttt{abdb}) = (\varphi_1, \varphi_2, \varphi_3, \varphi_4) = (2, 1, 0, 0)$, indicating that there are $\varphi_1 = 2$ symbols that appear once in `abdb` and $\varphi_2 = 1$ symbol that appears 2 times and so on. The sequences `abdb` and `dcca` for example have the same profile, though their patterns are different. The definition of a profile can be similarly extended to joint patterns or pairs of sequences and consists of entries φ_{μ_1,μ_2} that are the number of symbols that have appeared μ_1 times in first sequence and μ_2 times in the second sequence. For example,

$$\varphi(\texttt{dac}, \texttt{adbda}) = \varphi(123, 21412) = \begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 2 \end{array},$$

where the prevalances φ_{μ_1,μ_2} are arranged in a matrix with the rows indexed with μ_1 and columns with μ_2. As seen in the matrix, $\varphi_{1,2} = 2$, since there are 2 symbols, namely `d` and `a` that appear $\mu_1 = 1$ times in `dac` and $\mu_2 = 2$ times in `adbda`. By convention, we set $\varphi_{0,0} \equiv 0$.

Let $N(\varphi)$ be the number of patterns with the same profile φ and Φ^n be the set of all distinct profiles of sequences of length n. It was shown in [16] that the probability estimator for $\overline{\psi} \in \Psi^n$,

$$q(\overline{\psi}) \stackrel{\text{def}}{=} \frac{1}{|\Phi^n|}\frac{1}{N(\varphi(\overline{\psi}))},$$

which assigns equal probability estimate to all profiles and equal estimate to all patterns within a profile, is a good estimate for patten maximum likelihood, *i.e.,* $q(\overline{\psi}) \geq \hat{p}(\overline{\psi})\exp(-\pi\sqrt{2n/3})$.

We consider a similar estimator for maximum likelihood of joint patterns. Namely, denoting the number of joint patterns with the same profile φ by $N(\varphi)$ and the set of all distinct profiles of length-(n, n) sequences by $\Phi^{n,n}$, we consider the estimator

$$q_{\text{jp}}(\overline{\psi}_1, \overline{\psi}_2) \stackrel{\text{def}}{=} \frac{1}{|\Phi^{n,n}|}\frac{1}{N(\varphi(\overline{\psi}_1, \overline{\psi}_2))}$$

for joint patterns $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n,n}$.

We use the estimators q and q_{jp} instead of the pattern maximum likelihoods in $\Delta^{\hat{P}(\Psi)}$ and consider the test $\Delta^{N(\varphi)}_{n,\delta}$ defined as

$$\Delta^{N(\varphi)}_{n,\delta}(\overline{x}_1, \overline{x}_2) \stackrel{\text{def}}{=} \begin{cases} \textit{diff} & \text{if } \frac{N(\varphi(\overline{x}_1\overline{x}_2))}{N(\varphi(\overline{x}_1, \overline{x}_2))} > \frac{1}{\sqrt{\delta}}, \\ \textit{same} & \text{otherwise,} \end{cases}$$

for all $(\overline{x}_1, \overline{x}_2) \in \mathcal{A}^n \times \mathcal{A}^n$ and for some $\delta < \exp(-14n^{2/3})$. Our second main result, Theorem 12, shows that the test $\Delta^{N(\varphi)}$ also has low error probability

$$P_{\mathrm{e,sym}}(\Delta^{N(\varphi)}, p_1, p_2) \le 2\sqrt{\delta}\exp(7n^{2/3})$$

when (p_1, p_2) are either identical or (n, δ)-different. In the process, we show a convexity result for profile probabilities, that resembles the convexity of KL-divergence.

We note that $N(\varphi)$ can be calculated by the expressions

$$N(\varphi) = \frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!},$$

for $\varphi \in \Phi^n$ [16] and as shown in Appendix B,

$$N(\varphi) = \frac{(n!)^2}{\prod_{\mu_1,\mu_2}^{n} (\mu_1!\mu_2!)^{\varphi_{\mu_1,\mu_2}} \varphi_{\mu_1,\mu_2}!},$$

for $\varphi \in \Phi^{n,n}$. Hence, for $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n,n}$, the quantity $\frac{N(\varphi(\overline{\psi}_1\overline{\psi}_2))}{N(\varphi(\overline{\psi}_1,\overline{\psi}_2))}$ can be evaluated efficiently with time and space complexity both $\mathcal{O}(n)$.

We look at Example 1 again, this time using the test $\Delta^{N(\varphi)}$. When $(\overline{X}_1, \overline{X}_2) \sim (p_1, p_2)$ and $\Psi(\overline{X}_1, \overline{X}_2) = (\overline{\psi}_1, \overline{\psi}_2) = (1^n, 1^{\frac{n}{2}}23\cdots(\frac{n}{2}+1))$, the profile $\varphi = \varphi(\overline{\psi}_1, \overline{\psi}_2)$ has $\varphi_{0,1} = \frac{n}{2}$, $\varphi_{n,\frac{n}{2}} = 1$ and all other $\varphi_{\mu_1,\mu_2} = 0$. And the profile $\varphi' = \varphi(\overline{\psi}_1\overline{\psi}_2)$ has $\varphi'_1 = \frac{n}{2}$, $\varphi'_{\frac{3n}{2}} = 1$ and all other $\varphi'_\mu = 0$. Hence, by Stirling approximation,

$$\frac{N(\varphi(\overline{X}_1\overline{X}_2))}{N(\varphi(\overline{X}_1,\overline{X}_2))} = N(\varphi')/N(\varphi) = \frac{(2n)!}{(\frac{3n}{2})!\cdot(\frac{n}{2})!} \Big/ \frac{(n!)^2}{n!(\frac{n}{2})!\cdot(\frac{n}{2})!} \approx \left(\frac{4}{3}\right)^{\frac{3}{2n}} > 1.53^n,$$

and the test $\Delta_{n,\delta}^{N(\varphi)}$ outputs ***diff*** for a suitable δ as in the case of $\Delta^{\hat{P}(\Psi)}$, say $\delta = \exp(-16n^{2/3})$. When $(\overline{X}_1, \overline{X}_2) \sim (p_2, p_2)$, and $\Psi(\overline{X}_1, \overline{X}_2) = (\overline{\psi}_1, \overline{\psi}_2) = (1^{\frac{n}{2}}23\cdots a_{\frac{n}{2}+1}, 1^{\frac{n}{2}}(\frac{n}{2}+2)\cdots(n+1))$,

$$\frac{N(\varphi(\overline{X}_1\overline{X}_2))}{N(\varphi(\overline{X}_1,\overline{X}_2))} = \frac{(2n)!}{n!n!} \Big/ \frac{(n!)^2}{(\frac{n}{2})!(\frac{n}{2})!\cdot(\frac{n}{2})!\cdot(\frac{n}{2})!} \approx \frac{\sqrt{\pi n}}{2}$$

and the output of $\Delta_{n,\delta}^{N(\varphi)}$ is *same* for $\delta = \exp(-16n^{2/3})$.

While the error probability results that we show for the tests $\Delta_{n,\delta}^{N(\varphi)}$ and $\Delta_{n,\delta}^{\hat{P}(\Psi)}$ are useful only when $\delta < \exp(-14n^{2/3})$, for higher values of δ, we can characterize their performance in terms of *sample complexity*. It is shown in Corollary 14 that if (p_1, p_2) are (δ, n)-different for some $\delta < \frac{1}{4}$, then the test $\Delta_{n',\delta'}^{N(\varphi)}$ also has error probability less than δ when given sequences of length

$$n' = \max\left\{19n, \frac{120000n^3}{(\log_2 \frac{1}{4\delta})^3}\right\},$$

where $\delta' = \delta^2 \exp(-14n'^{2/3})$. In particular, if $\delta < \exp(-19n^{2/3})$, the error probability of $\Delta^{N(\varphi)}$ is less than δ when given $n' = 19n$ samples.

2 Error analysis of the test $\Delta^{\hat{P}(\Psi)}$

In order to analyze the error probability of $\Delta_{n,\delta}^{\hat{P}(\Psi)}$, we show some ancillary results on profiles of joint patterns and their probabilities.

We begin by showing bounds on $|\Phi^{n,n}|$, the number of profiles of joint patterns, and show that $|\Phi^{n,n}|$ is subexponential in the sequence length. To count the number of profiles $|\Phi^{n_1,n_2,\ldots,n_d}|$, we relate it to *partitions* of $(n_1, n_2, \ldots, n_d)$. We say that a multiset of d-tuples of non-negative integers $\{(\mu_{1,i}, \mu_{2,i}, \ldots, \mu_{d,i})\}_{i=1}^m$ is an (unordered) partition of $(n_1, n_2, \ldots, n_d)$ if $\sum_{i=1}^m \mu_{j,i} = n_j$ for $j = 1, 2, \ldots, d$. The sum of two d-tuples denotes their component-wise sum, *i.e.,* $(\mu_1, \mu_2, \ldots, \mu_d)$ + $(\mu'_1, \mu'_2, \ldots, \mu'_d) \overset{\mathrm{def}}{=} (\mu_1 + \mu'_1, \mu_2 + \mu'_2, \ldots, \mu_d + \mu'_d)$. The product of a scalar with a d-tuple is component-wise product with the scalar, *i.e.,* $\alpha \cdot (\mu_1, \mu_2, \ldots, \mu_d) \overset{\mathrm{def}}{=} (\alpha \cdot \mu_1, \alpha \cdot \mu_2, \ldots, \alpha \cdot \mu_d)$. Thus, $\{(0,1), (0,1), (2,1)\}$ is an unordered partition of $(2,3)$, because $2 \cdot (0,1) + (2,1) = (2,3)$.

We denote the number of partitions of $(n_1, n_2, \ldots, n_d)$ by the *joint partition function* $P(n_1, n_2, \ldots, n_d)$. For example, $P(2,1) = 4$, since

$$(2,1) = (1,0) + (1,1) = (2,0) + (0,1) = 2 \cdot (1,0) + (0,1).$$

Observation 1. *For all $d \geq 1$ and non-negative integers $n_1, n_2, \ldots, n_d$,*

$$|\Phi^{n_1,n_2,\ldots,n_d}| = P(n_1, n_2, \ldots, n_d).$$

□

It is a well known result due to Hardy and Ramanujan [8, 9] that for all n, the partition function $P(n)$ is bounded as

$$\exp\Big(\pi\sqrt{\frac{2}{3}}\sqrt{n}(1 - o(1))\Big) \leq P(n) < \exp\Big(\pi\sqrt{\frac{2}{3}}\sqrt{n}\Big).$$

The following lemma shows an upper bound on $P(n_1, n_2, \ldots, n_d)$, similar to [22, Theorem 15.7].

Lemma 2. *For all $d \geq 1$ and all $n_1, n_2, \ldots, n_d \geq 2^{d+1}$,*

$$P(n_1, n_2, \ldots, n_d) \leq \exp\Big(2\Big(1 + \frac{1}{d}\Big)\sum_{j=1}^{d} n_j^{d/(d+1)}\Big).$$

Proof. A proof is shown in Appendix A. □

Corollary 3. *For all $d \geq 1$ and $n \geq 2^{d+1}$,*

$$|\Phi^{n,n,\ldots\ (d\ \text{times})}| = P(n, n, \ldots\ (d\ \text{times})) < \exp\Big(2(d+1)n^{d/(d+1)}\Big).$$

□

Let $(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_d) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2} \times \cdots \mathcal{A}^{n_d}$ be generated *i.i.d.* and independently according to $(p_1, p_2, \ldots, p_d)$ respectively. The probability of a profile $\varphi \in \Phi^{n_1,n_2,\ldots,n_d}$ under $(p_1, p_2, \ldots, p_d)$ is the probability of observing a list of sequences with that profile, *i.e.,*

$$\begin{aligned} p_{1,2,\ldots,d}(\varphi) &\overset{\text{def}}{=} p_{1,2,\ldots,d}\Big(\varphi(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_d) = \varphi\Big) \\ &= \sum_{\substack{(\overline{\psi}_1,\overline{\psi}_2,\ldots,\overline{\psi}_d): \\ \varphi(\overline{\psi}_1,\overline{\psi}_2,\ldots,\overline{\psi}_d)=\varphi}} p_{1,2,\ldots,d}(\overline{\psi}_1, \overline{\psi}_2, \ldots, \overline{\psi}_d). \end{aligned}$$

Joint patterns with the same profile have the same probability when the sequences are generated by *i.i.d.* distributions. Hence, for all $(\overline{\psi}_1, \overline{\psi}_2, \ldots, \overline{\psi}_d)$,

$$p_{1,2,\ldots,d}(\varphi(\overline{\psi}_1, \overline{\psi}_2, \ldots, \overline{\psi}_d)) = N(\varphi(\overline{\psi}_1, \overline{\psi}_2, \ldots, \overline{\psi}_d)) \cdot p_{1,2,\ldots,d}(\overline{\psi}_1, \overline{\psi}_2, \ldots, \overline{\psi}_d).$$

The following lemma provides a simple bound on the probability of generating sequences whose profile has low probability.

Lemma 4. *Let $(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_d) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2} \times \cdots \times \mathcal{A}^{n_d}$ be generated* i.i.d. *according to $(p_1, p_2, \ldots, p_d)$ respectively, where $n_1, n_2, \ldots, n_d \geq 2^{d+1}$. Then, for all $0 < \delta \leq 1$,*

$$\Pr\Big(p_{1,2,\ldots,d}(\varphi(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_d)) < \delta\Big) < \delta \exp\Big(2\Big(1 + \frac{1}{d}\Big)\sum_{j=1}^{d} n_j^{d/(d+1)}\Big).$$

Proof. Using union bound and Corollary 3,

$$\begin{aligned} \Pr\Big(p_{1,2,\ldots,d}(\varphi(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_d)) < \delta\Big) &\leq \sum_{\varphi: p_{1,2,\ldots,d}(\varphi) < \delta} p_{1,2,\ldots,d}(\varphi) \\ &< \delta|\Phi^{n_1,n_2,\ldots,n_d}| \\ &\leq \delta \exp\Big(2\Big(1 + \frac{1}{d}\Big)\sum_{j=1}^{d} n_j^{d/(d+1)}\Big). \end{aligned}$$

Corollary 5. *Let $(\overline{X}_1, \overline{X}_2) \in \mathcal{A}^n \times \mathcal{A}^n$ be generated* i.i.d. *according to (p_1, p_2), where $n \geq 8$. Then, for all $0 < \delta \leq 1$,*

$$\Pr\Big(p_{1,2}(\varphi(\overline{X}_1, \overline{X}_2)) < \delta\Big) < \delta \exp(6n^{2/3}).$$

□

We make the following observation on (n, δ)-different distributions before we proceed to analyze the error probability of $\Delta^{\hat{P}(\Psi)}$.

Observation 6. *Let (p_1, p_2) be a pair of distributions over $\mathcal{A}$ that are (n, δ)-different. And let $\varphi \in \Phi^{n,n}$ be a profile such that $p_{1,2}(\varphi) \geq \delta$. Then, for all distributions p_3 over $\mathcal{A}$, $p_{3,3}(\varphi) < \delta$.*

Proof. Suppose on the contrary, there exists a distribution p_3 such that $p_{3,3}(\varphi) \geq \delta$. Any symmetric test Δ labels all sequence pairs having profile $\varphi \in \Phi^{n,n}$ as either ***same*** or ***diff***. If it maps them as ***same***, then $P_{\mathrm{e}}^n(\Delta, p_1, p_2) \geq \delta$ and if it maps them as ***diff***, then $P_{\mathrm{e}}^n(\Delta, p_3, p_3) \geq \delta$, *i.e.,* one of the error probabilities is $\geq \delta$, which contradicts the fact that (p_1, p_2) are (δ, n)-different. □

The following theorem shows that the test $\Delta_{n,\delta}^{\hat{P}(\Psi)}$ has low error probability both when the distribution pairs are same or (n, δ)-different.

Theorem 7. *For all $n \geq 8$, all $0 < \delta < \exp(-12n^{2/3})$, and all pairs distributions (p_1, p_2) that are either same or (n, δ)-different,*

$$P_{\mathrm{e}}^n(\Delta_{n,\delta}^{\hat{P}(\Psi)}, p_1, p_2) < \sqrt{\delta} \exp(6n^{2/3}).$$

Proof. Let $(\overline{X}_1, \overline{X}_2) \sim p_1^n \times p_2^n$. Consider the case when the (p_1, p_2) are same, *i.e.,* $p_1 = p_2$. Then,

$$\begin{aligned}
P_{\mathrm{e}}^n(\Delta^{\hat{P}(\Psi)}, p_1, p_1) &= \Pr\Big(\frac{\hat{p}(\Psi(\overline{X}_1, \overline{X}_2))}{\hat{p}(\Psi(\overline{X}_1 \overline{X}_2))} > \frac{1}{\sqrt{\delta}}\Big) \\
&\overset{\text{(a)}}{=} \Pr\Big(\frac{\hat{p}(\Psi(\overline{X}_1, \overline{X}_2))}{p_{3,3}(\Psi(\overline{X}_1, \overline{X}_2))} > \frac{1}{\sqrt{\delta}}\Big) \\
&\overset{\text{(b)}}{=} \Pr\Big(\frac{\hat{p}(\varphi(\overline{X}_1, \overline{X}_2))}{p_{3,3}(\varphi(\overline{X}_1, \overline{X}_2))} > \frac{1}{\sqrt{\delta}}\Big) \\
&\overset{\text{(c)}}{\leq} \Pr\Big(\frac{1}{p_{1,1}(\varphi(\overline{X}_1, \overline{X}_2))} > \frac{1}{\sqrt{\delta}}\Big) \\
&\overset{\text{(d)}}{<} \sqrt{\delta} \exp(6n^{2/3}),
\end{aligned}$$

where in (a), $p_3 = \arg\max_p p(\Psi(\overline{X}_1 \overline{X}_2))$ and in (b), we convert pattern probabilities to profile probabilities by multiplying and dividing by $N(\varphi(\overline{X}_1, \overline{X}_2))$ and using $p(\varphi) = N(\varphi) p(\overline{\psi}_1, \overline{\psi}_2)$. For (c), we use that $\hat{p}(\varphi(\overline{X}_1, \overline{X}_2)) \leq 1$ and we use Corollary 5 for (d).

Now consider the case when (p_1, p_2) are (n, δ)-different. For a sequence pair $(\overline{X}_1, \overline{X}_2)$, let $p_3 = \arg\max_p p(\Psi(\overline{X}_1 \overline{X}_2))$. Then,

$$\begin{aligned}
\Pr\Big(\frac{\hat{p}(\Psi(\overline{X}_1, \overline{X}_2))}{\hat{p}(\Psi(\overline{X}_1 \overline{X}_2))} \leq \frac{1}{\sqrt{\delta}}\Big) &\leq \Pr\Big(\frac{p_{1,2}(\Psi(\overline{X}_1, \overline{X}_2))}{p_{3,3}(\Psi(\overline{X}_1, \overline{X}_2))} \leq \frac{1}{\sqrt{\delta}}\Big) \\
&= \Pr\Big(\frac{p_{1,2}(\varphi(\overline{X}_1, \overline{X}_2))}{p_{3,3}(\varphi(\overline{X}_1, \overline{X}_2))} \leq \frac{1}{\sqrt{\delta}}\Big) \\
&< \sqrt{\delta} \exp(6n^{2/3}).
\end{aligned}$$

For the last step, in the case when $p_{1,2}(\varphi) \geq \sqrt{\delta}$, there is no error since Observation 6 implies that for all p_3, $p_{3,3}(\varphi) < \delta$ and hence $\frac{p_{1,2}(\varphi(\overline{X}_1, \overline{X}_2))}{p_{3,3}(\varphi(\overline{X}_1, \overline{X}_2))} > \frac{\sqrt{\delta}}{\delta} = \frac{1}{\sqrt{\delta}}$. Hence, the error probability is bounded by the probability of the case when $p_{1,2}(\varphi) < \sqrt{\delta}$, which by Corollary 5 is $< \sqrt{\delta} \exp(6n^{2/3})$.

As mentioned in Section 1, direct computation of maximum likelihood of patterns in the test $\Delta^{\hat{P}(\Psi)}$ may be difficult and hence we look at a computationally easier test $\Delta^{N(\varphi)}$.

3 Error analysis of the test $\Delta^{N(\varphi)}$

We show a few more useful results for analyzing the error probability of $\Delta^{N(\varphi)}$, which relate the quantities $N(\varphi)$, the number of patterns in a profile and $\hat{P}(\varphi)$, the maximum likelihood of the profile under *i.i.d.* distributions.

The *type* of a sequence $\overline{x} \in \mathcal{A}^n$ is the vector of multiplicities $\tau(\overline{x}) \overset{\text{def}}{=} \big(\mu(a_1), \mu(a_2), \ldots, \mu(a_k)\big)$, where $\mu(a_i)$ is the number of appearances of a_i in $\overline{x}$ for $i = 1, 2, \ldots, k$. Similarly, the *joint type* of a pair of sequences $(\overline{x}_1, \overline{x}_2) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$ is the vector of multiplicity pairs $\tau(\overline{x}_1, \overline{x}_2) \stackrel{\text{def}}{=} \big((\mu_1(a_1), \mu_2(a_1)), (\mu_1(a_2), \mu_2(a_2)), \ldots, (\mu_1(a_k), \mu_2(a_k))\big)$, where $\mu_1(a_i)$ and $\mu_2(a_i)$ are the number of appearances of a_i in $\overline{x}_1$ and $\overline{x}_2$ for $i = 1, 2, \ldots, k$. The set of all possible distinct types of sequences in $\mathcal{A}^n$ is denoted by $\mathcal{T}^n$ and the set of all possible distinct joint types of sequences in $\mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$ is denoted by $\mathcal{T}^{n_1,n_2}$.

The probability of a type $\tau = \big(\mu(a_i)\big)_{i=1}^{k} \in \mathcal{T}^n$ under a distribution p over $\mathcal{A}$ is

$$p(\tau) \stackrel{\text{def}}{=} \sum_{\tau(\overline{x})=\tau} p(\overline{x}) \;=\; \binom{n}{\mu(a_1), \mu(a_2), \cdots, \mu(a_k)} \prod_{i=1}^{k} p(a_i)^{\mu(a_i)},$$

i.e., the probability of observing a sequence whose type is τ. Similarly, the probability of a joint type $\tau = \big((\mu_1(a_i), \mu_2(a_i))\big)_{i=1}^{k} \in \mathcal{T}^{n_1,n_2}$ under a pair of distributions (p_1, p_2) over $\mathcal{A}$ is

$$\begin{aligned} p_{1,2}(\tau) &\stackrel{\text{def}}{=} \sum_{\tau(\overline{x}_1,\overline{x}_2)=\tau} p_{1,2}(\overline{x}_1, \overline{x}_2) \\ &= \binom{n_1}{\mu_1(a_1),\ \mu_1(a_2),\ \cdots,\ \mu_1(a_k)} \binom{n_2}{\mu_2(a_1),\ \mu_2(a_2),\ \cdots,\ \mu_2(a_k)} \prod_{i=1}^{k} p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu_2(a_i)}. \end{aligned}$$

The *sum type* of a joint type $\tau = \big((\mu_1(a_i), \mu_2(a_i))\big)_{i=1}^{k} \in \mathcal{T}^{n,n}$ is $\tau_s(\tau) \stackrel{\text{def}}{=} \big(\mu(a_i)\big)_{i=1}^{k} \in \mathcal{T}^{2n}$, where $\mu(a_i) \stackrel{\text{def}}{=} \mu_1(a_i) + \mu_2(a_i)$ for $i = 1, 2, \ldots, k$. The probability of a (sum) type $\tau_s \in \mathcal{T}^{2n}$ under a pair of distributions $p_{1,2} = (p_1, p_2)$ is the probability of the set of all types $\tau \in \mathcal{T}^{n,n}$ such that $\tau_s(\tau) = \tau_s$, *i.e.*,

$$p_{1,2}(\tau_s) \stackrel{\text{def}}{=} \sum_{\substack{\tau \in \mathcal{T}^{n,n}: \\ \tau_s(\tau)=\tau_s}} p_{1,2}(\tau).$$

For any pair of distributions (p_1, p_2) over $\mathcal{A} \times \mathcal{A}$, $p_{1/2} \stackrel{\text{def}}{=} (p_1 + p_2)/2$ denotes the distribution over $\mathcal{A}$ such that $p_{1/2}(a_i) = (p_1(a_i) + p_2(a_i))/2$ for $i = 1, 2, \ldots, k$.

Observation 8. *For all types $\tau_s \in \mathcal{T}^{2n}$ and all (p_1, p_2),*

$$\sum_{\substack{\tau \in \mathcal{T}^{n,n}: \\ \tau_s(\tau)=\tau_s}} p_{1,2}(\tau) \;=\; p_{1,2}(\tau_s) \;\le\; p_{1/2}(\tau_s) \frac{(n!)^2 2^{2n}}{(2n)!} \;<\; p_{1/2}(\tau_s) \sqrt{\pi n} e^{\frac{1}{6n}}.$$

Proof. Let $\tau_s = \big(\mu(a_i)\big)_{i=1}^{k}$. Then,

$$\begin{aligned} p_{1,2}(\tau_s) &= \sum_{\substack{\tau \in \mathcal{T}^{n,n}: \\ \tau_s(\tau)=\tau_s}} p_{1,2}(\tau) \\ &= \sum_{\substack{(\mu_1(a_1),\ldots,\mu_1(a_k)): \\ 0 \le \mu_1(a_i) \le \mu(a_i) \text{ for } i=1,\ldots,k, \\ \text{and } \mu_1(a_1)+\cdots+\mu_1(a_k)=n}} n!n! \prod_{i=1}^{k} \frac{1}{\mu_1(a_i)!(\mu(a_i) - \mu_1(a_i))!} p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu(a_i)-\mu_1(a_i)} \\ &= \frac{n!n!}{\prod_{i=1}^{k} \mu(a_i)!} \sum_{\substack{(\mu_1(a_1),\ldots,\mu_1(a_k)): \\ 0 \le \mu_1(a_i) \le \mu(a_i) \text{ for } i=1,\ldots,k, \\ \text{and } \mu_1(a_1)+\cdots+\mu_1(a_k)=n}} \prod_{i=1}^{k} \binom{\mu(a_i)}{\mu_1(a_i)} p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu(a_i)-\mu_1(a_i)} \\ &\le \frac{n!n!}{\prod_{i=1}^{k} \mu(a_i)!} \sum_{\substack{(\mu_1(a_1),\ldots,\mu_1(a_k)): \\ 0 \le \mu_1(a_i) \le \mu(a_i) \text{ for } i=1,\ldots,k}} \prod_{i=1}^{k} \binom{\mu(a_i)}{\mu_1(a_i)} p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu(a_i)-\mu_1(a_i)} \\ &= \frac{n!n!}{\prod_{i=1}^{k} \mu(a_i)!} \prod_{i=1}^{k} \Big(\sum_{\mu_1(a_i)=0}^{\mu(a_i)} \binom{\mu(a_i)}{\mu_1(a_i)} p_1(a_i)^{\mu_1(a_i)} p_2(a_i)^{\mu(a_i)-\mu_1(a_i)} \Big) \\ &= \frac{n!n!}{\prod_{i=1}^{k} \mu(a_i)!} \prod_{i=1}^{k} (p_1(a_i) + p_2(a_i))^{\mu(a_i)} \end{aligned}$$

$$= \frac{(n!)^2 2^{2n}}{(2n)!} \binom{2n}{\mu(a_1), \mu(a_2), \ldots, \mu(a_k)} \prod_{i=1}^{k} \Big(\frac{p_1(a_i) + p_2(a_i)}{2}\Big)^{\mu(a_i)}$$

$$= \frac{(n!)^2 2^{2n}}{(2n)!} p_{1/2}(\tau_s). \qquad \square$$

The profile of a type $\tau \in \mathcal{T}^n$ is $\varphi(\tau) = \varphi(\overline{x})$, where $\overline{x}$ is any sequence whose type is $\tau(\overline{x}) = \tau$. Similarly, for any $\tau \in \mathcal{T}^{n_1,n_2}$, $\varphi(\tau) \overset{\text{def}}{=} \varphi(\overline{x}_1, \overline{x}_2)$, where $(\overline{x}_1, \overline{x}_2)$ is any sequence pair such that $\tau(\overline{x}_1, \overline{x}_2) = \tau$.

Observation 9. *For all profiles $\varphi \in \Phi^n$ and all distributions p,*

$$p(\varphi) = \sum_{\varphi(\tau)=\varphi} p(\tau).$$

Likewise, for all profiles $\varphi \in \Phi^{n_1,n_2}$ and all pairs of distributions (p_1, p_2),

$$p_{1,2}(\varphi) = \sum_{\varphi(\tau)=\varphi} p_{1,2}(\tau). \qquad \square$$

The *sum profile* of a profile $\varphi \in \Phi^{n,n}$ is $\varphi_s(\varphi) \overset{\text{def}}{=} \varphi(\overline{\psi}_1 \overline{\psi}_2) \in \Phi^{2n}$ where $(\overline{\psi}_1, \overline{\psi}_2)$ is any joint pattern having profile $\varphi(\overline{\psi}_1, \overline{\psi}_2) = \varphi$. Hence, if $\varphi = [\varphi_{\mu_1,\mu_2}]$, where $\mu_1 = 0, 1, \ldots, n$ and $\mu_2 = 0, 1, \ldots, n$, then $\varphi_s = (\varphi'_1, \varphi'_2, \ldots, \varphi'_{2n})$ is given by $\varphi'_\mu = \sum_{\mu_1+\mu_2=\mu} \varphi_{\mu_1,\mu_2}$. The probability of a (sum) profile $\varphi_s \in \Phi^{2n}$ under a pair of distributions $p_{1,2}$ is the probability $p_{1,2}$ assigns to the set of all profiles $\varphi \in \Phi^{n,n}$ such that $\varphi_s(\varphi) = \varphi_s$, *i.e.*,

$$p_{1,2}(\varphi_s) \overset{\text{def}}{=} \sum_{\substack{\varphi \in \Phi^{n,n}: \\ \varphi_s(\varphi)=\varphi_s}} p_{1,2}(\varphi).$$

The following lemma on profile probabilities is analogous to the convexity of KL-divergence.

Lemma 10. *For all $\varphi_s \in \Phi^{2n}$ and all (p_1, p_2),*

$$\sum_{\substack{\varphi \in \Phi^{n,n}: \\ \varphi_s(\varphi)=\varphi_s}} p_{1,2}(\varphi) \;=\; p_{1,2}(\varphi_s) \;\le\; p_{1/2}(\varphi_s) \frac{(n!)^2 2^{2n}}{2n!} \;<\; p_{1/2}(\varphi_s) \sqrt{\pi n} e^{\frac{1}{6n}}.$$

Proof. Using Observations 8 and 9,

$$\begin{aligned} p_{1,2}(\varphi_s) &= \sum_{\substack{\varphi \in \Phi^{n,n}: \\ \varphi_s(\varphi)=\varphi_s}} p_{1,2}(\varphi) \\ &= \sum_{\varphi(\tau_s)=\varphi_s} p_{1,2}(\tau_s) \\ &\le \sum_{\varphi(\tau_s)=\varphi_s} \frac{(n!)^2 2^{2n}}{(2n)!} p_{1/2}(\tau_s) \\ &= p_{1/2}(\varphi_s) \frac{(n!)^2 2^{2n}}{(2n)!}. \end{aligned} \qquad \square$$

The following Lemma 11 relates the ratio of maximum likelihoods of any joint pattern $(\overline{\psi}_1, \overline{\psi}_2)$ and its concatenated pattern $\overline{\psi}_1 \overline{\psi}_2$ which appear in the test $\Delta^{\hat{P}(\Psi)}$, to the ratio of counts of patterns in their respective profiles, *i.e.*, $N(\varphi(\overline{\psi}_1, \overline{\psi}_2))$ and $N(\varphi(\overline{\psi}_1 \overline{\psi}_2))$ that appear in the test $\Delta^{N(\varphi)}$.

Lemma 11. *For all joint patterns $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n,n}$,*

$$\frac{N(\varphi(\overline{\psi}_1 \overline{\psi}_2))}{N(\varphi(\overline{\psi}_1, \overline{\psi}_2))} \;\ge\; \frac{\hat{p}(\overline{\psi}_1, \overline{\psi}_2)}{\hat{p}(\overline{\psi}_1 \overline{\psi}_2)} \frac{(2n)!}{(n!)^2 2^{2n}} \;>\; \frac{\hat{p}(\overline{\psi}_1, \overline{\psi}_2)}{\hat{p}(\overline{\psi}_1 \overline{\psi}_2)} \frac{1}{\sqrt{\pi n} e^{\frac{1}{6n}}}.$$

Proof. Let $p_{1,2} = (p_1, p_2)$ be such that $\hat{p}(\overline{\psi}_1, \overline{\psi}_2) = p_{1,2}(\overline{\psi}_1, \overline{\psi}_2)$. Note that $\varphi_s(\varphi(\overline{\psi}_1, \overline{\psi}_2)) = \varphi(\overline{\psi}_1\overline{\psi}_2)$. Using Lemma 10, we have

$$\begin{aligned}
N(\varphi(\overline{\psi}_1, \overline{\psi}_2))\hat{p}(\overline{\psi}_1, \overline{\psi}_2) &= N(\varphi(\overline{\psi}_1, \overline{\psi}_2))p_{1,2}(\overline{\psi}_1, \overline{\psi}_2) \\
&= p_{1,2}(\varphi(\overline{\psi}_1, \overline{\psi}_2)) \\
&\leq p_{1,2}(\varphi_s(\varphi(\overline{\psi}_1, \overline{\psi}_2))) \\
&\leq p_{1/2}(\varphi_s(\varphi(\overline{\psi}_1, \overline{\psi}_2)))\frac{(n!)^2 2^{2n}}{(2n)!} \\
&= p_{1/2}(\varphi(\overline{\psi}_1\overline{\psi}_2))\frac{(n!)^2 2^{2n}}{(2n)!} \\
&\leq \hat{p}(\varphi(\overline{\psi}_1\overline{\psi}_2))\frac{(n!)^2 2^{2n}}{(2n)!} \\
&= N(\varphi(\overline{\psi}_1\overline{\psi}_2))\hat{p}(\overline{\psi}_1\overline{\psi}_2)\frac{(n!)^2 2^{2n}}{(2n)!}.
\end{aligned}$$

□

Theorem 12. *For all $n \geq 8$, all $0 < \delta < \frac{1}{4\pi n e^{1/3n}}\exp(-12n^{2/3})$, and all pairs distributions (p_1, p_2) that are either same or (n, δ)-different,*

$$P_{\mathrm{e}}^n(\Delta^{N(\varphi)}, p_1, p_2) < \sqrt{\delta}\exp(6n^{2/3})\sqrt{\pi n}e^{\frac{1}{6n}}.$$

Proof. Let $(\overline{X}_1, \overline{X}_2) \sim p_1^n \times p_2^n$. Consider the case when $p_1 = p_2$. Then,

$$\begin{aligned}
P_{\mathrm{e}}^n(\Delta^{N(\varphi)}, p_1, p_1) &= \Pr\Big(\frac{N(\varphi(\overline{X}_1\overline{X}_2))}{N(\varphi(\overline{X}_1, \overline{X}_2))} > \frac{1}{\sqrt{\delta}}\Big) \\
&\overset{\text{(a)}}{=} \Pr\Big(\frac{p_1(\varphi(\overline{X}_1\overline{X}_2))}{p_{1,1}(\varphi(\overline{X}_1, \overline{X}_2))} > \frac{1}{\sqrt{\delta}}\Big) \\
&\leq \Pr\Big(\frac{1}{p_{1,1}(\varphi(\overline{X}_1, \overline{X}_2))} > \frac{1}{\sqrt{\delta}}\Big) \\
&< \sqrt{\delta}\exp(6n^{2/3}),
\end{aligned}$$

where in (a), we used $\frac{N(\varphi(\overline{\psi}_1\overline{\psi}_2))}{N(\varphi(\overline{\psi}_1,\overline{\psi}_2)} = \frac{N(\varphi(\overline{\psi}_1\overline{\psi}_2))p_1(\overline{\psi}_1\overline{\psi}_2)}{N(\varphi(\overline{\psi}_1,\overline{\psi}_2)p_{1,1}(\overline{\psi}_1,\overline{\psi}_2)} = \frac{p_1(\varphi(\overline{\psi}_1\overline{\psi}_2))}{p_{1,1}(\varphi(\overline{\psi}_1,\overline{\psi}_2))}$ and the last inequality is due to Corollary 5.

Consider the case when (p_1, p_2) are (n, δ)-different. For a sequence pair $(\overline{X}_1, \overline{X}_2)$, let $p_3 = \arg\max_p p(\Psi(\overline{X}_1\overline{X}_2))$. Then,

$$\begin{aligned}
P_{\mathrm{e}}^n(\Delta^{N(\varphi)}, p_1, p_2) &= \Pr\Big(\frac{N(\varphi(\overline{X}_1\overline{X}_2))}{N(\varphi(\overline{X}_1, \overline{X}_2))} \leq \frac{1}{\sqrt{\delta}}\Big) \\
&\overset{\text{(a)}}{\leq} \Pr\Big(\frac{1}{\sqrt{\pi n}e^{\frac{1}{6n}}}\frac{\hat{p}(\Psi(\overline{X}_1, \overline{X}_2))}{\hat{p}(\Psi(\overline{X}_1\overline{X}_2))} \leq \frac{1}{\sqrt{\delta}}\Big) \\
&\leq \Pr\Big(\frac{p_{1,2}(\Psi(\overline{X}_1, \overline{X}_2))}{p_{3,3}(\Psi(\overline{X}_1\overline{X}_2))} \leq \frac{\sqrt{\pi n}e^{\frac{1}{6n}}}{\sqrt{\delta}}\Big) \\
&< \sqrt{\delta}\exp(6n^{2/3})\sqrt{\pi n}e^{\frac{1}{6n}},
\end{aligned}$$

where in (a), we used Lemma 11 For the last inequality, we again consider the cases $p_{1,2}(\varphi(\overline{X}_1, \overline{X}_2)) \geq \sqrt{\delta}\sqrt{\pi n}e^{\frac{1}{6n}}$ and $< \sqrt{\delta}\sqrt{\pi n}e^{\frac{1}{6n}}$ separately similar to the proof of Theorem 12. In the case when $p_{1,2}(\varphi(\overline{X}_1, \overline{X}_2)) \geq \sqrt{\delta}\sqrt{\pi n}e^{\frac{1}{6n}} > \delta$, Observation 6 implies $p_{3,3}(\varphi(\overline{X}_1, \overline{X}_2)) < \delta$. Hence, $\frac{p_{1,2}(\varphi(\overline{X}_1,\overline{X}_2))}{p_{3,3}(\varphi(\overline{X}_1,\overline{X}_2))} > \frac{\sqrt{\delta}\sqrt{\pi n}e^{\frac{1}{6n}}}{\delta} = \frac{\sqrt{\pi n}e^{\frac{1}{6n}}}{\sqrt{\delta}}$ and hence, this case does not contribute to error probability. The error probability is therefore bounded by the probability of the other case, which by Corollary 5 is bounded as $\Pr\Big(p_{1,2}(\varphi(\overline{X}_1, \overline{X}_2)) < \sqrt{\delta}\sqrt{\pi n}e^{\frac{1}{6n}}\Big) < \sqrt{\delta}\exp(6n^{2/3})\sqrt{\pi n}e^{\frac{1}{6n}}$. □

4 Sample complexity of closeness testing

The error analysis results of Theorems 7 and 12 can be rephrased in terms of sample complexity. Also, Theorems 7 and 12 are applicable only when $\delta \le \exp(-14n^{2/3})$, and this section partially addresses the general case when $\delta < \frac{1}{2}$.

Observation 13. *If (p_1, p_2) are (n, δ)-different distributions for some $0 < \delta < \frac{1}{2}$, then they are also (n', δ')-different, where*

$$n' = \min\left\{20n, \frac{15000n^3}{D(\frac{1}{2}||\delta)^3}\right\} \text{ and } \delta' \le \delta^2 \exp(14n'^{2/3}),$$

where $D(\delta_1||\delta_2) \overset{\text{def}}{=} \delta_1 \log \frac{\delta_1}{\delta_2} + (1-\delta_1)\log\frac{1-\delta_1}{1-\delta_2}$.

Proof sketch. Since (p_1, p_2) are (n, δ)-different, for any p_3 there is a test that can distinguish (p_1, p_2) and (p_3, p_3) with error probability $< \delta$. We can obtain another test for sequences of length $n' = (2r+1)n$ such that the error probability of this test is $\delta' = \sum_{i=r+1}^{2r+1} \delta^i \binom{2r+1}{i}(1-\delta)^{2r+1-i}$ by using the original test on $(2r+1)$ pairs of length-n sequences and outputting the majority decision. It can be verified that $(2r+1) \ge \min\{19, \frac{15000n^2}{D(\frac{1}{2}||\delta)^3}\}$ suffices to guarantee that $\sum_{i=r+1}^{2r+1} \delta^i \binom{2r+1}{i}(1-\delta)^{2r+1-i} \le \delta^2 \exp(14((2r+1)n)^{2/3})$. □

Corollary 14. *If (p_1, p_2) are (n, δ)-different distributions for some $0 < \delta < \frac{1}{4}$, then they are also (n', δ')-different where $\delta' \le \delta^2 \exp(14n'^{2/3})$ for $n' = \max\left\{19n, \frac{120000n^3}{(\log_2 \frac{1}{4\delta})^3}\right\}$. Furthermore if $\delta < \exp(-19n^{2/3})$, then $n' = 19n$ suffices.* □

Hence, using Theorem 12 and Corollary 14, it follows that whenever (p_1, p_2) are identical or (n, δ)-different, the error probability of the test $\Delta^{N(\varphi)}_{n',\delta'}$ is less than δ, using sequences of length $n' = \max\left\{19n, \frac{120000n^3}{(\log_2 \frac{1}{4\delta})^3}\right\}$, where $\delta' = \delta^2 \exp(14n'^{2/3})$.

5 Related problems and open problems

For the problem of classification described in 1.3, our results imply that whenever the distributions of the classes, p_1 and p_2, are (n, δ)-different, the closeness tests $\Delta^{\hat{P}(\Psi)}$ or $\Delta^{N(\varphi)}$ can be used to construct classifiers whose error probability is $\le 2\sqrt{\delta}\exp(7n^{2/3})$. We define two distributions (p_1, p_2) to be (n, δ)-classifiable if length-n sequence pairs generated by (p_1, p_2) can be distinguished with error probability $< \delta$ from those generated by (p_1, p_1) and (p_2, p_2) by a symmetric test. While (n, δ)-difference implies (n, δ)-classifiabilty, it remains to answer if the opposite is also true.

As mentioned earlier, our results are applicable when the error probabilities (δ) are small and $\le \exp(-14n^{2/3})$, and while we partially address the case of general $\delta < \frac{1}{2}$, and it remains to perform a better analysis. We also hope to reduce the subexponential factor of $\exp(7n^{2/3})$ in the right hand side of Theorems 7 and 12 using a tighter analysis.

Lastly, it remains to fully answer the question of when two distributions (p_1, p_2) are (n, δ)-different. In many cases such as Example 1, the quantity $\frac{N(\varphi(\overline{X}_1\overline{X}_2))}{N(\varphi(\overline{X}_1,\overline{X}_2))}$ in the test $\Delta^{N(\varphi)}$ can be shown to be exponentially large in n with high probability, that implies (n, δ)-difference for a suitable δ. This question is also answered in part by [3] and [21] where distributions are parametrized in terms of alphabet size.

References

[1] J. Acharya, H. Das, A. Orlitsky, S. Pan, and N.P. Santhanam. Classification using pattern probability estimators. In *Proceedings of IEEE Symposium on Information Theory*, pages 1493–1497, 2010.

[2] T. Batu, L. Fortnow, E. Fischer, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. *FOCS '01: Proceedings of the 42nd Annual Symposium on Foundations of Computer Science*, page 442, 2001.

[3] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 259, Washington, DC, USA, 2000. IEEE Computer Society.

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

[5] Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, page 2005, 2002.

[6] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley Interscience, 2nd edition, 2006.

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

[8] G.H. Hardy and S. Ramanujan. Asymptotic formulae in combinatory analysis. *Proceedings of London Mathematics Society*, 17(2):75–115, 1918.

[9] G.H. Hardy and E.M. Wright. *An introduction to the theory of numbers.* Oxford University Press, 1985.

[10] 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.

[11] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses.* Springer Texts in Statistics. Springer, New York, third edition, 2005.

[12] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[13] A. Orlitsky and Shengjun Pan. The maximum likelihood probability of skewed patterns. In *Proceedings of IEEE Symposium on Information Theory*, pages 1130–1134, 2009.

[14] A. Orlitsky and N.P. Santhanam. Speaking of infinity. *IEEE Transactions on Information Theory*, 50:2215–2230, 2004.

[15] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Limit results on pattern entropy. *IEEE Transactions on Information Theory*, 52:2954–2964, 2006.

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

[17] H. V. Poor. *An introduction to signal detection and estimation.* New York: Springer-Verlag, 2nd edition, 1994.

[18] Sofya Raskhodnikova. *Property Testing: Theory and Applications.* PhD thesis, Massachusetts Institute of Technology, 2003.

[19] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. In *FOCS '07: Proceedings of the 48th Annual Symposium on Foundations of Computer Science*, 2007.

[20] N.P. Santhanam, A. Orlitsky, and K. Viswanathan. New tricks for old dogs: Large alphabet probability estimation. In *Information Theory Worskshop*, pages 638–643, 2007.

[21] Paul Valiant. Testing symmetric properties of distributions. In *STOC '08: Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 383–392, New York, NY, USA, 2008. ACM.

[22] J.H. van Lint and R.M. Wilson. *A course in combinatorics.* Cambridge University Press, 2001.

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

A Number of profiles of a given length

Lemma 15. *For all $d \geq 1$ and all $n_1, n_2, \ldots, n_d \geq 2^{d+1}$,*

$$P(n_1, n_2, \ldots, n_d) \leq \exp\left(2\Big(1 + \frac{1}{d}\Big)\sum_{j=1}^{d} n_j^{d/(d+1)}\right).$$

Proof. The (ordinary) generating function of $P(n_1, n_2, \ldots, n_d)$ is

$$G(x_1, x_2, \ldots, x_d) = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\cdots\sum_{n_d=0}^{\infty} P(n_1, n_2, \ldots, n_d) x_1^{n_1} x_2^{n_2} \cdots x_d^{n_d} = \prod_{\substack{(\mu_1,\mu_2,\ldots,\mu_d)\\ \in\mathbb{N}^d\setminus(0,0,\ldots,0)}} \frac{1}{1 - x_1^{\mu_1} x_2^{\mu_2} \cdots x_d^{\mu_d}},$$

where $\mathbb{N} = \{0, 1, 2, \cdots\}$ and $0 < x_1, x_2, \ldots, x_d < 1$. Hence,

$$\begin{aligned}
\log G(x_1, x_2, \ldots, x_d) &= \sum_{\substack{(\mu_1,\mu_2,\ldots,\mu_d)\\ \in\mathbb{N}^d\setminus(0,0,\ldots,0)}} -\log\Big(1 - \prod_{j=1}^{d} x_j^{\mu_j}\Big)\\
&= \sum_{\substack{(\mu_1,\mu_2,\ldots,\mu_d)\\ \in\mathbb{N}^d\setminus(0,0,\ldots,0)}} \sum_{l=1}^{\infty} \frac{1}{l}\Big(\prod_{j=1}^{d} x_j^{\mu_j}\Big)^l\\
&= \sum_{l=1}^{\infty}\frac{1}{l} \sum_{\substack{(\mu_1,\mu_2,\ldots,\mu_d)\\ \in\mathbb{N}^d\setminus(0,0,\ldots,0)}} \prod_{j=1}^{d} (x_j^l)^{\mu_j}\\
&= \sum_{l=1}^{\infty}\frac{1}{l}\left(\frac{1}{\prod_{j=1}^{d}(1 - x_j^l)} - 1\right)\\
&= \sum_{l=1}^{\infty}\frac{1}{l}\frac{1 - \prod_{j=1}^{d}(1 - x_j^l)}{\prod_{j=1}^{d}\Big((1 - x_j)\Big(\sum_{i=0}^{l-1} x_j^i\Big)\Big)}\\
&< \sum_{l=1}^{\infty}\frac{1}{l}\frac{1 - \prod_{j=1}^{d}(1 - x_j^l)}{\Big(\prod_{j=1}^{d}(1 - x_j)\Big)\Big(1 + \sum_{j=1}^{d}\sum_{i=1}^{l-1} x_j^i\Big)}\\
&\overset{\text{(a)}}{<} \frac{1}{\prod_{j=1}^{d}(1 - x_j)}\left(1 + \sum_{l=2}^{\infty}\frac{1}{l(l-1)}\right)\\
&= \frac{2}{\prod_{j=1}^{d}(1 - x_j)}.
\end{aligned}$$

In the Inequality (a), we consider the cases $l = 1$ and $l > 1$ separately. When $l > 1$, in the denominator, $\Big(1 + \sum_{j=1}^{d}\sum_{i=1}^{l-1} x_j^i\Big) > (l - 1)\sum_{j=1}^{d} x_j^i > (l - 1)\Big(1 - \prod_{j=1}^{d}(1 - x_j^l)\Big)$. Since $G(x_1, x_2, \ldots, x_d) > P(n_1, n_2, \ldots, n_d) x^{n_1} x^{n_2} \cdots x^{n_d}$, we have

$$\log P(n_1, n_2, \ldots, n_d) \ < \ \log G(x_1, x_2, \ldots, x_d) - \sum_{j=1}^{d} n_j \log x_j \ < \ \frac{2}{\prod_{j=1}^{d}(1 - x_j)} - \sum_{j=1}^{d} n_j \log x_j.$$

Substituting $x_j = 1 - n_j^{-1/(d+1)}$ for $j = 1, 2, \ldots, d$, we get

$$\log P(n_1, n_2, \ldots, n_d) \ < \ 2\prod_{j=1}^{d} n_j^{1/(d+1)} + \sum_{j=1}^{d} n_j \log\Big(1 - n_j^{-1/(d+1)}\Big) \ \leq \ 2\Big(1 + \frac{1}{d}\Big)\sum_{j=1}^{d} n_j^{d/(d+1)}.$$

In the last step, we used AM-GM inequality, *i.e.,* $\prod_{j=1}^{d} n_j^{1/(d+1)} = \big(\prod_{j=1}^{d} n_j^{d/(d+1)}\big)^{1/d} \leq \frac{1}{d}\sum_{j=1}^{d} n_j^{d/(d+1)}$, and $\log(1 - \epsilon) < 2\epsilon$ for $\epsilon \leq \frac{1}{2}$, hence $\log\big(1 - n_j^{-1/(d+1)}\big) \leq 2n_j^{-1/(d+1)}$ for $n_j > 2^{d+1}$ and $j = 1, 2, \ldots, d$. □

B Number of patterns of a given profile

The number of joint patterns with the same profile φ is denoted by$N(\varphi)$. For example, consider the profile $\varphi = \varphi(1232, 13)$ which has $\varphi_{1,1} = 2$, $\varphi_{2,0} = 1$ and all other $\varphi_{\mu_1,\mu_2} = 0$. Then, $N(\varphi) = 12$ since the set of all joint patterns that have this profile is $\{(1123, 23), (1123, 32), (1213, 23), (1213, 32), (1223, 13), (1223, 31), (1231, 23), (1231, 32), (1232, 13), (1232, 31), (1233, 13), (1233, 21)\}$. The following lemma gives an expression for $N(\varphi)$ and extends Lemma 3 in [16].

Lemma 16. *For all $d \geq 1$ and all $\varphi \in \Phi^{n_1,n_2,\ldots,n_d}$,*

$$N(\varphi) = \frac{\prod_{j=1}^{d} n_d!}{\prod_{\mu_1=0}^{n_1}\prod_{\mu_2=0}^{n_2}\cdots\prod_{\mu_d=0}^{n_d}(\mu_1!\mu_2!\cdots\mu_d!)^{\varphi_{\mu_1,\mu_2,\ldots,\mu_d}}\varphi_{\mu_1,\mu_2,\ldots,\mu_d}!}.$$

Proof. We show the lemma for $d = 2$, and the proof is similar for any $d \geq 1$. Let $\varphi \in \Phi^{n_1,n_2}$. Any joint pattern $(\overline{\psi}_1, \overline{\psi}_2)$ that has profile φ is a pair of sequences with symbols from $\{1, 2, \ldots, m\}$, where $m = \sum_{\mu_1=0}^{n_1}\sum_{\mu_2=0}^{n_2}\varphi_{\mu_1,\mu_2}$ is the total number of symbols in $\overline{\psi}_1\overline{\psi}_2$. Let $\{\mu_1(i)\}_{i=1}^m$ and $\{\mu_2(i)\}_{i=1}^m$ be non-negative integers such that $\sum_{i=1}^m \mu_1(i) = n_1$ and $\sum_{i=1}^m \mu_2(i) = n_2$. The number of sequence pairs whose alphabet is $\{1, 2, \ldots, m\}$, and the number of appearances of i in first sequence is $\mu_1(i)$ and in second sequence is $\mu_2(i)$, for $i = 1, 2, \ldots, m$, is

$$\binom{n_1}{\mu_1(1), \mu_1(2), \ldots, \mu_1(m)}\binom{n_2}{\mu_2(1), \mu_2(2), \ldots, \mu_2(m)} = \frac{n_1!n_2!}{\prod_{i=1}^{m}\mu_1(i)!\mu_2(i)!}.$$

The number of different ways of choosing $\{\mu_1(i)\}_{i=1}^m$ and $\{\mu_2(i)\}_{i=1}^m$ such it conforms to profile is φ is

$$\binom{m}{\varphi_{0,0},\ \varphi_{0,1},\ \ldots,\ \varphi_{n_1,n_2}} = \frac{m!}{\prod_{\mu_1=0}^{n_1}\prod_{\mu_2=0}^{n_2}\varphi_{\mu_1,\mu_2}!}.$$

Thus, the number of sequence pairs whose alphabet is $\{1, 2, \ldots, m\}$ and profile is φ is

$$N^*(\varphi) = \frac{n_1!n_2!}{\prod_{i=1}^{m}\mu_1(i)!\mu_2(i)!}\frac{m!}{\prod_{\mu_1=0}^{n_1}\prod_{\mu_2=0}^{n_2}\varphi_{\mu_1,\mu_2}!} = \frac{n_1!n_2!m!}{\prod_{\mu_1=0}^{n_1}\prod_{\mu_2=0}^{n_2}(\mu_1!\mu_2!)^{\varphi_{\mu_1,\mu_2}}\varphi_{\mu_1,\mu_2}!}.$$

Clearly, $N^*(\varphi) = m! \cdot N(\varphi)$, since

- $\geq$: For each joint pattern having profile φ, the labels $\{1, 2, \ldots, m\}$ can be permuted in $m!$ ways to generate $m!$ different sequence pairs whose alphabet is $\{1, 2, \ldots, m\}$ and profile is φ. Furthermore, the sets of sequence pairs generated in this way by different joint patterns are disjoint. So $N^*(\varphi) \geq m! \cdot N(\varphi)$.

- $\leq$: Given any pair of sequences $(\overline{x}_1, \overline{x}_2)$ having alphabet $\{1, 2, \ldots, m\}$ and profile φ, their symbols can be permuted keeping the positions same to obtain a joint pattern with profile φ, which is in fact $\Psi(\overline{x}_1, \overline{x}_2)$. There are exactly $m!$ sequence pairs having alphabet $\{1, 2, \ldots, m\}$ and profile φ that have the same joint pattern. Hence, $N^*(\varphi) \leq m! \cdot N(\varphi)$.

Thus,

$$N(\varphi) = \frac{N^*(\varphi)}{m!} = \frac{n_1!n_2!}{\prod_{\mu_1=0}^{n_1}\prod_{\mu_2=0}^{n_2}(\mu_1!\mu_2!)^{\varphi_{\mu_1,\mu_2}}\varphi_{\mu_1,\mu_2}!}.$$

□

C Symmetric tests

We provide a formal treatment to the intuition that joint patterns of sequences contain sufficient information for the problem of closeness testing, similar to [4],[3],[21].

We define the *symmetric error probability* of a test Δ for (p_1, p_2) as its worst case error probability over all possible permutations of the alphabet, *i.e.,*

$$P^n_{\text{e,sym}}(\Delta, p_1, p_2) \stackrel{\text{def}}{=} \max_{\sigma \in S_{\mathcal{A}}} P^n_{\text{e}}(\Delta, p_1^\sigma, p_2^\sigma),$$

where $S_{\mathcal{A}}$ is the set of all permutations of $\mathcal{A}$. Clearly, since separation between distributions does not depend on the actual symbols, and depends only the probability multiset, it is appropriate to look at the symmetric error probability.

A *symmetric* test is a test whose output does not change when the alphabet is permuted and gives the same output for all sequence pairs which have the same joint pattern, *i.e.,* $\Delta(\overline{x}_1, \overline{x}_2) = \tilde{\Delta}(\Psi(\overline{x}_1, \overline{x}_2))$ for all $(\overline{x}_1, \overline{x}_2)$, where $\tilde{\Delta} : \Psi^{n,n} \to \{\texttt{same}, \texttt{diff}\}$. Hence, a symmetric test depends only the joint pattern of the sequences. Note that for a symmetric test Δ, $P_{\mathrm{e,sym}}(\Delta, p_1, p_2) = P_{\mathrm{e}}(\Delta, p_1, p_2)$ for all distribution pairs (p_1, p_2). The following observation shows that we may limit ourselves to considering only symmetric closeness tests.

Observation 17. *Let $\Delta : \mathcal{A}^n \times \mathcal{A}^n \to \{\texttt{same}, \texttt{diff}\}$ be any test for closeness, possibly not symmetric. Then, there exists a symmetric test $\tilde{\Delta} : \mathcal{A}^n \times \mathcal{A}^n \to \{\texttt{same}, \texttt{diff}\}$ such that for all pairs of distributions (p_1, p_2) over $\mathcal{A}$, $P^n_{\mathrm{e,sym}}(\tilde{\Delta}, p_1, p_2) \le 2 \cdot P^n_{\mathrm{e,sym}}(\Delta, p_1, p_2)$.*

Proof. Let $\tilde{\Delta}$ be the test whose output for a sequence pair is same as that made by Δ for the majority of sequence pairs with the same joint pattern, *i.e.,* $\tilde{\Delta}(\overline{x}_1, \overline{x}_2) = \text{majority}\{\Delta(\overline{x}'_1, \overline{x}'_2) : \Psi(\overline{x}'_1, \overline{x}'_2) = \Psi(\overline{x}_1, \overline{x}_2)\}$. Clearly, $P^n_{\mathrm{e}}(\tilde{\Delta}, p^{\sigma}_{1,2})$ is same for all permutations σ of $\mathcal{A}$. Thus, if p_1, p_2 are similar,

$$\begin{aligned}
P^n_{\mathrm{e,sym}}(\tilde{\Delta}, p_1, p_2) &= P^n_{\mathrm{e}}(\tilde{\Delta}, p_1, p_2)\\
&= \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} P^n_{\mathrm{e}}(\tilde{\Delta}, p^{\sigma}_{1,2})\\
&= \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} \sum_{\substack{(\overline{x}_1, \overline{x}_2):\\ \tilde{\Delta}(\overline{x}_1, \overline{x}_2) = \texttt{diff}}} p^{\sigma}_1(\overline{x}_1) p^{\sigma}_2(\overline{x}_2)\\
&= \sum_{\substack{(\overline{x}_1, \overline{x}_2):\\ \tilde{\Delta}(\overline{x}_1, \overline{x}_2) = \texttt{diff}}} \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} p^{\sigma}_1(\overline{x}_1) p^{\sigma}_2(\overline{x}_2)\\
&\overset{(a)}{\le} 2 \sum_{\substack{(\overline{x}_1, \overline{x}_2):\\ \Delta(\overline{x}_1, \overline{x}_2) = \texttt{diff}}} \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} p^{\sigma}_1(\overline{x}_1) p^{\sigma}_2(\overline{x}_2)\\
&= 2 \cdot \frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} \sum_{\substack{(\overline{x}_1, \overline{x}_2):\\ \Delta(\overline{x}_1, \overline{x}_2) = \texttt{diff}}} p^{\sigma}_1(\overline{x}_1) p^{\sigma}_2(\overline{x}_2)\\
&\le 2 \cdot \max_{\sigma \in S_{\mathcal{A}}} \sum_{\substack{(\overline{x}_1, \overline{x}_2):\\ \Delta(\overline{x}_1, \overline{x}_2) = \texttt{diff}}} p^{\sigma}_1(\overline{x}_1) p^{\sigma}_2(\overline{x}_2)\\
&= 2 \cdot P^n_{\mathrm{e,sym}}(\Delta, p_1, p_2),
\end{aligned}$$

where in (a), we note that all $(\overline{x}_1, \overline{x}_2)$ having the same joint pattern have the same probability $\frac{1}{|\mathcal{A}|!} \sum_{\sigma \in S_{\mathcal{A}}} p^{\sigma}_1(\overline{x}_1) p^{\sigma}_2(\overline{x}_2)$. A similar argument can be shown for the case $p_1 \ne p_2$. □