UNIVERSITY OF CALIFORNIA, SAN DIEGO

Competitive Tests and Estimators for Properties of Distributions

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Communication Theory and Systems)

by

Hirakendu Das

Committee in charge:

Professor Alon Orlitsky, Chair
Professor Ery Arias-Castro
Professor Sanjoy Dasgupta
Professor Young-Han Kim
Professor Paul Siegel

2012

The dissertation of Hirakendu Das is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2012

DEDICATION

To my family.

EPIGRAPH

What has been seen cannot be unseen.

— Unknown

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I am grateful to my advisor, colleagues, friends and family for their support and contribution in the making of this dissertation. My sincerest thanks go to my advisor Alon Orlitsky for the deep and fruitful interaction we had over the years. His unending list of qualities have always amazed and inspired me. In particular, I firmly believe that if there is a person who thinks differently, it is Alon. His deep, yet simple insights on many of the research problems we have worked on have been very helpful. In addition, I have always admired his approachability, frugality, and never ending pursuit for elegant solutions to both research and recreational problems. I am very thankful to my committee members Ery Arias-Castro, Sanjoy Dasgupta, Young-Han Kim and Paul Siegel for taking active interest in my research. I fondly remember attending the graduate courses taught by Young-Han Kim and Paul Siegel. I have also enjoyed taking many other courses at UCSD, especially those taught by Larry Milstein, Ramamohan Paturi and Alex Vardy.

I thank my colleagues Jayadev Acharya, Ashkan Jafarpour, Shengjun Pan, Ananda Theertha Suresh for the many technical and non-technical discussions. I am also thankful to Alon's former students and postdocs, Anand Dhulipala, Narayana Prasad Santhanam and Anand Sarwate, with whom I had brief but enlightening discussions. There is a long list of friends at UCSD and outside, which cannot fit into this page. I am happy and thankful to have come across them and to have enjoyed their company during the time of my PhD.

I am indebted to my parents and to my sister Sucheta, for their tireless love and support. They have done everything in their capacity, and beyond, to help me achieve whatever I have. Simply put, I am nothing without them.

Chapter 3 is adapted from Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, “Competitive closeness testing”, *Journal of Machine Learning Research - Proceedings Track (COLT 2011)*, 2011.

Chapter 4 is adapted from Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, Narayana Prasad Santhanam, “Classification using pattern maximum likelihood”, *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 2010.

Chapter 5 is adapted from Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, “Algebraic computation of pattern maximum likelihood”, *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 2011; and Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, Ananda Theertha Suresh, “Competitive estimation of discrete probability distributions”, In Preparation, 2012.

Chapter 6 is adapted from Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, “Estimating multisets of Bernoulli processes”, Submitted to *IEEE Symposium on Information Theory (ISIT)*, 2012; and Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, “Estimating multiple processes”, In preparation, 2012.

The dissertation author is a primary researcher and author of all of the above papers.

VITA

2006	B. Tech. in Electrical Engineering, Indian Institute of Technology Madras
2006-2012	Graduate Student Researcher, University of California, San Diego
2008	M. S. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego
2012	Ph. D. in Electrical Engineering (Communication Theory and Systems), University of California, San Diego

PUBLICATIONS

Hirakendu Das, Alon Orlitsky, Narayana Prasad Santhanam, Junan Zhang, “Further results on relative redundancy”, *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 1940 -1943, 2008.

Hirakendu Das, Alexander Vardy, “Multiplicity assignments for algebraic soft decoding”, *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 1248-1252, 2009.

Jayadev Acharya, Hirakendu Das, Olgica Milenkovic, Alon Orlitsky, Shengjun Pan, “String reconstruction using substring compositions”, *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 1238-1242, 2010.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, Narayana Prasad Santhanam, “Classification using pattern maximum likelihood”, *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 1493-1497, 2010.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, “Exact calculation of pattern probabilities”, *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 1498-1502, 2010.

Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, “Competitive closeness testing”, *Journal of Machine Learning Research - Proceedings Track (COLT 2011)*, 47-68, 2011.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, “Algebraic computation of pattern maximum likelihood”, *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 400-404, 2011.

Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, "Estimating multisets of Bernoulli processes", Submitted to *IEEE Symposium on Information Theory (ISIT)*, 2012.

Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, Ananda Theertha Suresh, "Competitive classification and closeness testing", Submitted to *Conference on Learning Theory (COLT)*, 2012.

Jayadev Acharya, Hirakendu Das, Alon Orlitsky, "On redundancy and distinguishability of label-invariant distributions", Submitted to *Conference on Learning Theory (COLT)*, 2012.

Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, Ananda Theertha Suresh, "Competitive estimation of discrete probability distributions", In Preparation, 2012.

ABSTRACT OF THE DISSERTATION

Competitive Tests and Estimators for Properties of Distributions

by

Hirakendu Das

Doctor of Philosophy in Electrical Engineering (Communication Theory and Systems)

University of California, San Diego, 2012

Professor Alon Orlitsky, Chair

We derive competitive tests and estimators for several properties of discrete distributions, based on their *i.i.d.* sequences. We focus on symmetric properties that depend only on the multiset of probability values in the distributions and not on specific symbols of the alphabet that assume these values. Many applications of probability estimation, statistics and machine learning involve such properties.

Our method of probability estimation, called profile maximum likelihood (PML), involves maximizing the likelihood of observing the profile of the given sequences, *i.e.,* the multiset of symbol counts in the sequences. It has been used successfully for universal compression of large alphabet data sources, and has been shown empirically to perform well for other probability estimation problems like classification and distribution multiset estimation. We provide competitive estimation guarantees for the PML method for several such problems.

For testing closeness of distributions, *i.e.,* finding whether two given *i.i.d.* sequences of length n are generated by the same distribution or by two different ones, our schemes have an error probability of at most $\sqrt{\delta} \cdot e^{7n^{2/3}}$ whenever the best possible error probability is $\delta \leq e^{-14n^{2/3}}$. The running time of our scheme is $\mathcal{O}(n)$. We do not make any assumptions on the distributions, including on their support size. In terms of sample complexity, this implies that if there is a closeness test which takes sequences of length n and has error probability at most δ, our tests have the same error guarantee on sequences of length $n' = \mathcal{O}\big(\max\{\frac{n^3}{(\log \frac{1}{4\delta})^3}, n\}\big)$. Similar results are implied for the related problem of classification.

For finding the probability multiset of a discrete distribution using a length-n *i.i.d.* sequence drawn from it, we show the following guarantee for the PML-based estimator. For any class of distributions and any distance metric on their probability multisets, if there is an estimator that approximates all distributions in this class to within a distance of $\epsilon > 0$ with error probability at most $\delta \leq e^{-6n^{1/2}}$, then the PML estimator is within a distance of 2ϵ with error probability at most $\delta \cdot e^{6n^{1/2}}$. Equivalently, the PML estimator approximates distributions to within a distance of 2ϵ with error probability δ using sequences of length $n' = \mathcal{O}\big(\max\{\frac{n^2}{(\log \frac{1}{4\delta})^2}, n\}\big)$. Thus, this estimator is competitive with other estimators, including the one by Valiant *et al.* [68] that approximates distributions of *superlinear* support size $k = \mathcal{O}(\epsilon^{2.1} n \log(n))$ to within a relative earthmover distance of ϵ and whose error probability can be shown to be at most $e^{-n^{0.9}}$. However, unlike the case of closeness testing, we do not yet have efficient schemes for computing the PML distribution.

We extend the results for PML for distribution multiset estimation to two related problems of estimating the parameter multiset of multiple distributions or processes. These include the problems of estimating the multiset of success probabilities of Bernoulli processes, and the multiset of means of Poisson distributions.

Chapter 1

Introduction

The increasing use of computers and internet means that more data, *e.g.,* text, pictures, music, video, is being produced than ever before. Despite faster computers and more efficient algorithms for analyzing and inferring from this large amount of data, there is much room for improvement, especially from the algorithms side. In most such data, the underlying alphabet is large, *e.g.,* the number of possible different words encountered in text data. Furthermore, the underlying distributions that produce this data often have a long "tail" and are therefore difficult to estimate. These problems have received a great amount of attention recently from researchers in the fields of machine learning, information theory, statistics, probability estimation, and more recently and prominently, from the property testing community in computer science. The work in this thesis aims to be a part of the larger effort for studying large alphabet data.

Many discrete data sources can be modeled as being independent and identically distributed (*i.i.d.*), *i.e.,* the atoms or symbols of the data are independent samples from an unknown discrete distribution. Although this is arguably one of the simplest models for probability estimation, our understanding of many basic inference tasks in this model is somewhat unsatisfactory, especially when it involves large alphabet distributions. We consider several problems of testing and estimating properties of distributions, given *i.i.d.* sequences generated by them. Our focus is on symmetric properties that depend only on the multiset of probability values in the distributions and not on the specific symbols of the alphabet that assume these values. For example, the entropy and L_2 norm of a distribution are symmetric properties, since they do not depend on how the probability values of the distribution are mapped to the alphabet. Similarly, the L_1 distance between two distributions is an example of a symmetric property of two distributions.

It is natural that for testing or estimating symmetric properties of distributions from *i.i.d.* sequences, we should rely on the *profile* of the given sequences, also known as *fingerprint* or *histogram of histograms*, that conveys the *multiset* of counts of various symbols in them, without referring to the symbols themselves. The profile of a sequence is different from *type*, that also conveys the counts of symbols in the sequence, but attaches these counts to the symbols' identities. Thus, the sequences `abac` and `cabb` have the same profile, since the multiset of symbol counts in both sequences is $\{2, 1, 1\}$, but not the same type, since the count of the symbol `a` in the two sequences is different.

The main technique we use for our tests and estimators is that of *profile maximum likelihood* (PML). As the name suggests, we consider the distribution that maximizes the likelihood of observing the given profile, *i.e.,* the profile of the given sequence, as the estimate of the underlying distribution. In general, this distribution is different from the one obtained by maximizing the sequence likelihood or equivalently type likelihood, which simply yields the naive estimate of empirical distribution, the normalized counts of symbols in the sequence. This technique, also known as *pattern maximum likelihood*[*], has been used by Orlitsky *et al.* in [49, 48] for probability estimation in the context of universal compression of large alphabet data sources. Motivated by this success and the "maximum likelihood principle" in general, it is natural to expect that this method would perform well for other problems of probability multiset estimation as well. Indeed, empirical results have shown that these techniques work very well for many instances of distribution estimation applications, *e.g.,* estimating the support size, total probability of unseen symbols (also known as "the missing mass problem"), properties related to the shape of the underlying distribution in [50, 75] and for classification of text documents in [58, 4].

[*] We defer the definition and motivation of patterns to the next chapter, and will observe that when considering *i.i.d.* sequences, maximum likelihood of profiles and patterns are equivalent

The results in this thesis provide theoretical guarantees for several such problems and show that the error performance of PML based methods is almost as good as that of any other test or estimator. In the process, we show several simple but powerful properties of maximum likelihood in general. These properties are of a competitive flavor and can be stated informally as follows.

Suppose there is a class of distributions, all of which are on a common discrete alphabet $\mathcal{Z}$. Given a sample in $\mathcal{Z}$ produced by one of these distributions, we want to estimate or test a property of these distributions. If there is an estimator or tester for this property for all distributions in this class such that its error probability is at most δ, then the error probability of the maximum likelihood tester or estimator is at most $\delta \cdot |\mathcal{Z}|$.

We treat the profile of the given sequences as our sample or observation, and thus, $\mathcal{Z}$ is the set of all profiles, whose size is shown to be subexponential: $|\mathcal{Z}| \leq e^{3n^{1/2}}$ for profiles of length-n sequences and $|\mathcal{Z}| \leq e^{6n^{2/3}}$ for profiles of pairs of length-n sequences. The class of distributions under consideration are those induced on profiles by *i.i.d.* sequences. As is often the case, a small constant distance of $\epsilon > 0$ in some metric between two distributions results in their profile distributions being highly separated and distinguishable with exponentially or near exponentially small error probabilities, implying the existence of estimators and testers with small error probabilities, *e.g.,* $\delta \leq e^{-n^{0.9}}$. As an easy example, also discussed in later chapters, if the L_1 distance between two distributions of support size $k = \mathcal{O}(\epsilon^{2.1} n)$ is $\epsilon > 0$, then their profiles, *i.e.,* the profiles of the length-n *i.i.d.* sequences they produce, are distinguishable with error probability $\leq e^{-n\epsilon^2/8}$. We note that this technique is similar to the well known *method of types*, *e.g.,* see [18, 19, 15], where $\mathcal{Z}$ can be considered as the set of all length-n types on alphabet of size k, whose size $|\mathcal{Z}| = \binom{n+k-1}{k-1}$ is polynomial in n, *i.e.,* $\mathcal{O}(n^k)$, when k is small compared to n.

A technical tool which we often find useful for analyzing distributions of profiles of *i.i.d.* sequences, and for developing estimators is the well known technique of “Poissonization”, *e.g.,* see [63, 7]. It relies mainly on two facts. The counts of various symbols are distributed independently in an *i.i.d.* sequence of length n', where n' is Poisson distribution with mean n, denoted by $\mathrm{poi}(n)$. Secondly, the Poisson distribution sharply concentrates around its mean. Using these facts, it can be shown that obtaining good distribution estimation guarantees using length-n *i.i.d.* sequences is almost equivalent to obtaining guarantees using *i.i.d.* sequences of length $n' \sim \mathrm{poi}(n)$, *e.g.,* see [68, 69]. Thus, it is convenient to analyze and solve estimation problems given sequences of length $\mathrm{poi}(n)$ since the independent symbol counts are often advantageous, while still implying similar results when given length-n sequences.

1.1 Closeness testing and classification

The first problem we study is that of testing closeness between two distributions. Given two length-n *i.i.d.* sequences from two unknown distributions, we want to test whether the distributions are same or different. There is an extensive amount of literature on this problem and several of its variants in the framework of hypothesis testing [44, 77, 30, 38, 15], which primarily considers asymptotic error performance when the sequence lengths tend to infinity. In such scenarios, it follows by Chernoff bounds that the empirical distributions of the sequences can be used as good estimates for the underlying distributions. Such tests can be shown to have low error probability when the alphabet size is $k = o(n)$. When alphabet size $k = \Omega(n)$, it is easy to show examples where two sequences generated by same distribution have very different empirical distributions with high probability.

For larger alphabets, Batu et al [8] developed a closeness test that distinguishes pairs of distributions that are same from those whose L_1 distance is at least $\epsilon > 0$, with error probability δ and using sequences of length n, whenever the alphabet size of the distributions is $k = \mathcal{O}\big(n^{3/2} \cdot \frac{\epsilon^4}{\log \frac{n}{\delta}}\big)$. They also show matching upper bounds that there exist two pairs of distributions on an alphabet of size $k = n^{3/2}$, one of which is a same pair and the other pair has a L_1 distance of 1, such that the distribution of the profiles of the sequences they generate are almost identical and hence cannot be distinguished with error probability less than $\frac{1}{4}$. Similar results have been shown for other distances like f-divergences by Guha *et al.* in [29]. Paul Valiant in [70] showed upper bounds on alphabet size for a related problem, that distinguishing between distribution pairs whose L_1 distance is $\leq \alpha$ from those distance is $\geq \beta$, for some $0 < \alpha < \beta < 2$, may be performed on all such distributions of support size at most k with low error probability, only when $k = n \cdot 2^{\mathcal{O}(\sqrt{\log(n)})}$. This upper bound has been improved by Valiant *et al.* in [69], along with a *linear* estimator that finds the L_1 distance between two distributions using two length-n sequences within an additive error of ϵ and low error probability, whenever $k = \mathcal{O}(\epsilon^2 n \log(n))$. A common strategy used in all these works is to rely on the empirical estimate, *i.e.,* normalized symbol counts, for high probability symbols, whereas the multiset of low probabilities symbols or their contribution to distances is estimated reliably by other statistics that concentrate, *e.g.,* collisions/coincidences in [7, 29] and profiles in [69]. For the related problem of classification on large alphabets, Kelly *et al.* in [35] show tests for distinguishing between distributions that are separated by a constant L_1 distance of $\epsilon > 0$ whose error probability is low whenever the $k = o(n^2)$ and all probabilities are $\Theta(\frac{1}{n^\alpha}) = \Theta(\frac{1}{k})$, for some $\alpha < 2$. We note that it is easy to see that there cannot be a test for all distributions of alphabet size $k = \Theta(n^2)$ by the Birthday problem.

Most of these previous works consider closeness testing in terms of a suitable distance measure and characterize the minimum number of samples required n as a function of the alphabet size k. These results have been equivalently stated above as characterization of the maximum range of alphabet size k in terms of n, for which all distributions can be tested for closeness, given sequences of length n. Clearly, the applicability of these results is restricted by upper bounds on k and moreover the algorithms require prior knowledge of k. For example, in spite of the example shown in [7] where a particular pair of distributions whose alphabet size is $k = n^{3/2}$ cannot be tested for closeness, there are many distribution pairs that have $k \gg n^{3/2}$ but can be distinguished trivially using sequences of length n. On a similar note, the separation in distances like L_1 together with the alphabet size bounds may not accurately reflect the number of samples required to distinguish these distributions. For example, consider a pair of distributions on a large alphabet k, one of which is a singleton, and the other has probability 1/2 on the same symbol as first one, and remaining mass of 1/2 equally spread over the other $k-1$ symbols. The L_1 distance of this pair is 1, but can be distinguished from all same pairs of distributions with error probability $\delta > 0$ whenever n is bigger than a constant multiple of $\log(\frac{1}{\delta})$, unlike the limiting example in [7] that requires sample sequences of length $n = \Omega(k^{2/3})$.

Instead we consider a notion of distance between distributions that is more natural for the problem of closeness testing. Informally, we say that a pair of distributions (n, δ)-different if it can be distinguished from all pairs of same distributions (possibly using different tests for various same pairs) using sequence pairs of length n and error probability δ. Note that following [7, 9], we only consider *symmetric* tests that depend on the pattern of the sequence pairs, and perform well regardless of how the symbols of the alphabet are mapped to the probability multiset of the two distributions. Clearly, the distribution pairs that are (n, δ)-different are the only ones we can possibly hope to distinguish from same pairs of distributions with error probability δ, using a single common test on length-n sequence pairs. We show closeness tests that that can distinguish between all same pairs of distributions and all pairs of (n, δ)-different distributions using sequence pairs of length-n and error probability $\sqrt{\delta} \cdot e^{7n^{2/3}}$. Thus, our tests are near optimal when $\delta \leq e^{-16n^{2/3}}$. Stated in terms of sample complexity, our tests can distinguish between (n, δ)-different and same pairs of distributions using sequences of length $n' = \mathcal{O}\big(\max\{\frac{n^3}{(\log \frac{1}{4\delta})^3}, n\}\big)$. Most importantly, we do not make any assumptions on the alphabet size or shape of the underlying pair of distributions, other than the fact that they are either same or (n, δ)-different.

Our tests have a rather simple and well known form. We note that the closeness testing problem and any property testing problem in general can be considered as what is known in statistics literature as composite hypothesis testing, *e.g.,* see [44, 54], for which the likelihood ratio test (LRT) is a commonly used method. Applied to closeness testing, and considering the profiles as the observations, our tests are LRTs on profiles. They compare the maximum likelihood of the profile of the given sequences under all same pairs of distributions to that under all possible pairs of distributions. If the ratio between these two maximum likelihoods is not too small, *i.e.,* $\geq e^{-7n^{2/3}}$, we declare the distribution pairs to be “same”, else we declare them “different”. The error guarantees for this test, follow from a similar result applicable to LRTs in general.

Since the computation of profile maximum likelihoods appears to be difficult in general, *e.g.,* see [50, 3], instead of directly using them in LRTs, we use one of their known approximations in terms of a combinatorial quantity which we refer to as the pattern counts of profiles and thus obtain a computationally efficient test. This test also offers similar performance guarantees.

In Chapter 3, we consider the closeness testing problem in detail. Implications to other property testing problems like testing uniformity are also discussed briefly. In Chapter 4, we show an application of our closeness testing results for classification. We also consider several direct approaches to this problem and show experimental results for various classifiers for text categorization.

1.2 Distribution multiset estimation and related problems

The next problem we study is that of estimating the probability multiset of a discrete distribution given an *i.i.d.* sequence of length n generated by it. Distribution estimation, both in terms of its multiset and the probabilities of specific observed symbols, has been studied for a long time. It dates back to the early famous works of Laplace [36], Fisher [24] and that of Good and Turing during World War II [26, 28, 27], followed by a long line of work by many researchers for a wide variety of applications, most prominently for finding abundance of species and language modeling [33, 25, 14, 61], and for the general problems of estimating probability of unseen elements [57, 34, 41, 42] and number of distinct elements in the underlying distribution [66, 13, 60, 74]. See Bunge and Fitzpatrick [11] for an overview of many applications and different approaches to this problem. Other recent interesting approaches to this problem include that of Jedynak and Khudanpur [32] and Wagner *et al.* [72]. Most of these works make certain assumptions on the underlying distributions, *e.g.,* a given prior or a restricted class of possible underlying distributions.

Moving away from such assumptions, there have been a number of recent works by the "property testing" community about distribution multiset estimation, which arises in the context of estimating *symmetric* properties of one or several distributions. Before looking at some of these results, we note that the empirical distribution is a good estimate of the underlying distribution with respect to most distances and suffices for accurately determining many properties when the alphabet size k is smaller than the sample sequence length n, *i.e.,* $k = o(n)$. Also, if we want the complete distribution, *i.e.,* the probabilities of specific symbols and not just the multiset, we cannot hope to estimate all distributions of support size k when $k = \Omega(n)$, *e.g.,* see [52]. The estimation guarantees that follow, hold with high probability over the given *i.i.d.* sequence. Batu *et al.* [9, 10] showed that the entropy of all distributions whose support size is $k = \widetilde{\mathcal{O}}(n^{(1-\epsilon)\gamma^2})$ can be estimated to within a factor of $\gamma > 1$ using a length-n sample sequence, for arbitrarily small $\epsilon > 0$. They also showed a similar *upper bound* (equivalently, lower bound on sample complexity) of $k = o(n^{2\gamma^2})$, by using the simple but illustrative example of uniform distributions on $n^{2\gamma^2}$ and n^2 symbols. For estimating the support size (number of probabilities $\geq \frac{1}{k}$) to within an additive error of ϵk, while it is easy to do so when $k = o(n)$, Raskhodnikova *et al.* [55] showed upper bounds of $k = o(n^{1+o(1)})$. Paul Valiant in [70] improved the upper bounds for γ-multiplicative approximation of entropy to $k = o(n^{\gamma^2})$ and an upper bound of $k = n \cdot 2^{\Theta(\sqrt{\log(n)})}$ for ϵk-additive approximation of support size. Valiant *et al.* in [68] show estimators for approximating the distribution multiset to within a *relative earthmover distance* of ϵ, and hence for ϵ-additive approximation of entropy and ϵk-additive approximation of support size, when $k = \mathcal{O}(\epsilon^{2.1} n \log(n))$. They also show matching upper bounds of $k = o(\epsilon^2 n \log(n))$. Augmenting these results, Valiant *et al.* in [68] show linear estimators of the form $\sum_{\mu=1}^{n} \alpha_\mu \varphi_\mu$ (where the α_μ's depend only on n and ϵ) for entropy and support size with similar estimation guarantees, and is related to the entropy estimator in [51]. Related results are also shown by Paninski, *e.g.,* for approximating entropy [51], approximating distributions in KL-divergence [52] and for testing uniformity [53].

We show that the profile maximum likelihood distribution, *i.e.,* the one that maximizes the probability of observing the same profile as that of the observed sequence, is as good as any other estimator in the following sense. For any class of distributions and any distance metric on their probability multisets, if there is an estimator that approximates all distributions in this class to within a distance of $\epsilon > 0$ with error probability at most $\delta \le e^{-6n^{1/2}}$, then our estimator is within a distance of 2ϵ with error probability at most $\delta \cdot e^{6n^{1/2}}$. Similar to the case of closeness testing, we show this via a general result about maximum likelihood for distribution estimation. In terms of sample complexity, the estimator approximates distributions to within a distance of 2ϵ with error probability δ using sequences of length $n' = \mathcal{O}\big(\max\{\frac{n^2}{(\log\frac{1}{4\delta})^2}, n\}\big)$. As an application of this result, we consider the estimator by Valiant *et al.* in [68] that approximates distributions of support size $k = \mathcal{O}(\epsilon^{2.1} n \log(n))$ to within a relative earthmover distance of ϵ and whose error probability is $e^{-n^{0.9}}$.

Despite the attractive estimation properties of the PML distribution, computing it efficiently appears to be difficult in general. We consider exact calculation of PML for short patterns using elimination methods from algebra. We also consider alternative approaches to distribution estimation, that are motivated by both computational efficiency and other criteria for measuring the quality of distribution estimates. Distribution multiset estimation is studied in Chapter 5.

In Chapter 6, we consider two problems of estimating the parameter multiset of distributions from a parametric family, which are related to the distribution multiset estimation problem. They are the problems of estimating the multiset of success probabilities of multiple independent Bernoulli processes, and that of estimating the multiset of means of a product of Poissons. We show that the PML estimator for the respective problems are as good as any other estimator. Furthermore, good distribution estimators can be used to construct good estimators for these problems and vice versa. The Poissonization technique mentioned previously is used at various points to connect the different problems.

Chapter 2

Preliminaries

2.1 Standard notation

We use the following standard notation throughout the thesis. For any positive integer z, the set $[z] \overset{\text{def}}{=} \{1, 2, \ldots, z\}$. For any set $\mathcal{Z}$, its size or cardinality is denoted by $|\mathcal{Z}|$. For any set $\mathcal{Z}$ and nonnegative integer n, unless defined otherwise, $\mathcal{Z}^n$ denotes the cross product $\mathcal{Z} \times \mathcal{Z} \times \cdots$ (n times), and $\mathcal{Z}^* \overset{\text{def}}{=} \cup_{n=0}^{\infty} \mathcal{Z}^n$.

For two integers $k \geq m \geq 0$, $[k]^{\underline{m}}$ is the set of all mappings (or permutations) $\sigma : [m] \to [k]$ and $k^{\underline{m}} \overset{\text{def}}{=} |[k]^{\underline{m}}| = k \cdot (k-1) \cdots (k-m+1) = \frac{k!}{(k-m)!}$ is the falling power m of k. In particular $S_k \overset{\text{def}}{=} [k]^{\underline{k}}$ is the symmetric group consisting of all permutations $\sigma : [k] \to [k]$ and $|S_k| = k!$.

2.2 Sequences, types and their likelihoods

Let $\mathcal{A} \overset{\text{def}}{=} \{a_1, a_2, \ldots, a_k\}$ denote a discrete alphabet of size $k \overset{\text{def}}{=} |\mathcal{A}|$. Let $\overline{x} \overset{\text{def}}{=} x_1 x_2 \ldots x_n$ be a sequence of length n with symbols in $\mathcal{A}$. The set $\mathcal{A}^n$ consists of all length-n sequences on $\mathcal{A}$ and $\mathcal{A}^*$ consists of all sequences.

The count or number of appearances of a symbol $a \in \mathcal{A}$ in $\overline{x}$ is

$$\mu(a) \overset{\text{def}}{=} \mu_{\overline{x}}(a) \overset{\text{def}}{=} |\{i : x_i = a\}| = \sum_{i=1}^{n} \mathbb{1}_{[x_i = a]},$$

and is also called the *multiplicity* of a. The vector of counts or multiplicities is

$$\mu(\overline{x}) \overset{\text{def}}{=} (\mu(a_1), \mu(a_2), \ldots, \mu(a_k)).$$

The *type* or empirical distribution of $\overline{x}$ is

$$\tau(\overline{x}) \overset{\text{def}}{=} \frac{\mu(\overline{x})}{n} \overset{\text{def}}{=} \Big(\frac{\mu(a_1)}{n}, \frac{\mu(a_2)}{n}, \ldots, \frac{\mu(a_k)}{n}\Big).$$

As such, the count vector and type convey the same information and are often used synonymously. In general, for any list of k nonnegative integers $\mu(a_1), \ldots, \mu(a_k)$ such that $\sum_{i=1}^{k} \mu(a_i) = n$, vectors $\overline{\mu} = (\mu(a_1), \ldots, \mu(a_k))$ and $\overline{\tau} = \frac{\overline{\mu}}{n}$ are a valid count vector and type of some length-n sequence respectively. The number of sequences with the same type $\overline{\tau}$ is

$$N(\overline{\tau}) \overset{\text{def}}{=} |\{\overline{x} : \tau(\overline{x}) = \overline{\tau}\}| = \binom{n}{\mu(a_1), \mu(a_2), \ldots, \mu(a_k)}$$

The set of all possible different types of length-n sequences, also called n-types, is denoted by

$$\mathcal{T}^n \overset{\text{def}}{=} \big\{\overline{\tau} : \overline{\tau} = \frac{\overline{\mu}}{n} \text{ and } \sum_{i=1}^{k} \mu(a_i) = n\big\}.$$

Hence, by a well known combinatorial fact, the number of distinct n-types is

$$|\mathcal{T}^n| = \binom{n+k-1}{k-1}.$$

Let $P \overset{\text{def}}{=} (P(a_1), P(a_2), \ldots, P(a_k))$ be a probability distribution on $\mathcal{A}$. Let $\overline{X} \overset{\text{def}}{=} X_1 X_2 \cdots X_n$ be a sequence of n random variables (r.v.'s) drawn *i.i.d.* according to P. We also say that $\overline{X}$ is a length-n *i.i.d.* sequence generated by P and $\overline{X} \sim P^n$ with the same meaning. Then, for all $\overline{x} \in \mathcal{A}^n$, the likelihood of $\overline{x}$ is

$$P(\overline{x}) \overset{\text{def}}{=} P^n(\overline{x}) \overset{\text{def}}{=} P(\overline{X} = \overline{x}) \overset{\text{def}}{=} \Pr\{\overline{X} = \overline{x}\} = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{k} P(a_i)^{\mu(a_i)},$$

the probability of observing $\overline{x}$ when $\overline{X} \sim P^n$. The distribution P^n on $\mathcal{A}^n$ is the distribution of $\overline{X}$.

The *sequence maximum likelihood* of $\overline{x}$ is its maximum likelihood (ML) under all distributions on $\mathcal{A}$ and denoted by

$$\hat{P}(\overline{x}) \overset{\text{def}}{=} \max_{P} P(\overline{x}).$$

The maximizing distribution is denoted by $\hat{P}_{\overline{x}} \overset{\text{def}}{=} \arg\max_P P(\overline{x})$ and it is an easy and well known fact that it is the empirical distribution, *i.e.,*

$$\hat{P}_{\overline{x}} = \tau(\overline{x}),$$

and therefore

$$\hat{P}(\overline{x}) = \hat{P}_{\overline{x}}(\overline{x}) = \prod_{i=1}^{k} \Big(\frac{\mu(a_i)}{n}\Big)^{\mu(a_i)}.$$

The probability or likelihood of a type $\overline{\tau} \in \mathcal{T}^n$ under P is

$$\begin{aligned} P(\overline{\tau}) \overset{\text{def}}{=} P(\tau(\overline{X}) = \overline{\tau}) &= \sum_{\overline{x}:\tau(\overline{x})=\overline{\tau}} P(\overline{x}) \\ &= N(\overline{\tau}) \prod_{i=1}^{k} P(a_i)^{\mu(a_i)}. \end{aligned}$$

Similar to maximum likelihood of sequences $\overline{x}$, we can define maximum likelihood of types $\overline{\tau}$, but the maximizing distribution is trivially the same when $\overline{\tau} = \tau(\overline{x})$.

2.2.1 Sequence tuples and joint types

Pairs and tuples of sequences are considered for closeness testing and problems to testing and estimating properties of multiple distributions. We mostly consider pairs of sequences here, although the definitions are easily extended to tuples. For pairs of sequences $(\overline{x}_1, \overline{x}_2) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$ whose lengths are n_1 and n_2, their vector of count pairs is

$$\mu(\overline{x}_1, \overline{x}_2) \overset{\text{def}}{=} \big((\mu_{\overline{x}_1}(a_1), \mu_{\overline{x}_2}(a_1)), \ldots, (\mu_{\overline{x}_1}(a_k), \mu_{\overline{x}_2}(a_k))\big),$$

and their joint type is

$$\tau(\overline{x}_1, \overline{x}_2) \overset{\text{def}}{=} \Big(\big(\frac{\mu_{\overline{x}_1}(a_1)}{n_1}, \frac{\mu_{\overline{x}_2}(a_1)}{n_2}\big), \ldots, \big(\frac{\mu_{\overline{x}_1}(a_k)}{n_1}, \frac{\mu_{\overline{x}_2}(a_k)}{n_2}\big)\Big).$$

We note that the definition of joint-type used here is different from that used usually in information theory, *e.g.,* [15], but is more natural for our problem since we consider sequence pairs or tuples that are generated *independently* (and *i.i.d.*) by different distributions (and thus conveys the same information as the usual definition in information theory).

Let $\overline{\mu}_1 = (\mu_1(a_1), \dots, \mu_1(a_k))$ and $\overline{\mu}_2 = (\mu_2(a_1), \dots, \mu_2(a_k))$ be any two count vectors such that $\sum_{i=1}^{k} \mu_1(a_i) = n_1$ and $\sum_{i=1}^{k} \mu_2(a_i) = n_2$. Then, the associated vector of pairs $\overline{\overline{\tau}} \stackrel{\text{def}}{=} \big((\frac{\mu_1(a_1)}{n_1}, \frac{\mu_2(a_1)}{n_2}), \dots (\frac{\mu_1(a_k)}{n_1}, \frac{\mu_2(a_k)}{n_2})\big)$ is a valid joint type. Let $\overline{\tau}_1$ and $\overline{\tau}_2$ be the projections, *i.e.,* first and second components of $\overline{\overline{\tau}}$, and therefore a valid n_1-type and n_2-type respectively. The number of sequence pairs with joint type $\overline{\overline{\tau}}$ is

$$\begin{aligned} N(\overline{\overline{\tau}}) &\stackrel{\text{def}}{=} |\{(\overline{x}_1, \overline{x}_2) : \tau(\overline{x}_1, \overline{x}_2) = \overline{\overline{\tau}}\}| = N(\overline{\tau}_1) \cdot N(\overline{\tau}_2) \\ &= \binom{n_1}{\mu_1(a_1), \dots, \mu_1(a_k)} \binom{n_2}{\mu_2(a_1), \dots, \mu_2(a_k)}. \end{aligned}$$

The set of all possible joint types of sequences of length (n_1, n_2) is denoted by $\mathcal{T}^{n_1,n_2} = \mathcal{T}^{n_1} \times \mathcal{T}^{n_2}$. Thus,

$$|\mathcal{T}^{n_1,n_2}| = |\mathcal{T}^{n_1}| \cdot |\mathcal{T}^{n_2}| = \binom{n_1+k-1}{k-1}\binom{n_2+k-1}{k-1}.$$

Let (P_1, P_2) be a pair of distributions on $\mathcal{A}$, *i.e.,* $P_1 \stackrel{\text{def}}{=} (P_1(a_1), \dots, P_1(a_k))$ and $P_2 \stackrel{\text{def}}{=} (P_2(a_1), \dots, P_2(a_k))$. Let $\overline{X}_1 \stackrel{\text{def}}{=} X_{1,1} \cdots X_{1,n}$ and $\overline{X}_2 \stackrel{\text{def}}{=} X_{2,1} \cdots X_{2,n}$ be two sequences drawn *i.i.d.* according to P_1 and P_2 respectively, *i.e.,* $\overline{X}_1 \sim P_1^{n_1}$ and $\overline{X}_2 \sim P_2^{n_2}$, also denoted as $(\overline{X}_1, \overline{X}_2) \sim P_1^{n_1} \times P_2^{n_2}$. For all $(\overline{x}_1, \overline{x}_2) \in \mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$, its likelihood under (P_1, P_2) is

$$\begin{aligned} P_{1,2}(\overline{x}_1, \overline{x}_2) &\stackrel{\text{def}}{=} \Pr\{(\overline{X}_1, \overline{X}_2) = (\overline{x}_1, \overline{x}_2)\} \\ &= P_1(\overline{x}_1)P_2(\overline{x}_2) = \prod_{i=1}^{k} P_1(a_i)^{\mu_1(a_i)} P_2(a_i)^{\mu_2(a_i)}, \end{aligned}$$

the probability of observing $\overline{x}$ when $\overline{X} \sim P^n$. The distribution $P_1^{n_1} \times P_2^{n_2}$ on $\mathcal{A}^{n_1} \times \mathcal{A}^{n_2}$ is simply the distribution of $(\overline{X}_1, \overline{X}_2)$.

The maximum likelihood of $(\overline{x}_1, \overline{x}_2)$ is

$$\hat{P}_{1,2}(\overline{x}_1, \overline{x}_2) \stackrel{\text{def}}{=} \max_{P_1, P_2} P_{1,2}(\overline{x}_1, \overline{x}_2).$$

The maximizing distribution is $\hat{P}_{\overline{x}_1, \overline{x}_2} \stackrel{\text{def}}{=} \arg\max_{P_1, P_2} P_{1,2}(\overline{x}_1, \overline{x}_2)$. It is easy to see that $\hat{P}_{\overline{x}_1, \overline{x}_2} = (\hat{P}_{\overline{x}_1}, \hat{P}_{\overline{x}_1}) = (\tau(\overline{x}_1), \tau(\overline{x}_2))$ and

$$\hat{P}_{1,2}(\overline{x}_1, \overline{x}_2) = \hat{P}_{\overline{x}_1}(\overline{x}_1)\hat{P}_{\overline{x}_2}(\overline{x}_2) = \prod_{i=1}^{k} \Big(\frac{\mu_1(a_i)}{n_2}\Big)^{\mu_1(a_i)} \Big(\frac{\mu_2(a_i)}{n_2}\Big)^{\mu_2(a_i)}.$$

The probability or likelihood of a joint type $\overline{\overline{\tau}} \in \mathcal{T}^{n_1,n_2}$ under (P_1, P_2) is

$$\begin{aligned} P_{1,2}(\overline{\overline{\tau}}) &\overset{\text{def}}{=} P_{1,2}(\tau(\overline{X}_1, \overline{X}_2) = \overline{\tau}) = \sum_{\overline{x}_1,\overline{x}_2:\tau(\overline{x}_1,\overline{x}_2)=\overline{\overline{\tau}}} P_1(\overline{x}_1)P_2(\overline{x}_2) \\ &= N(\overline{\tau}_1)N(\overline{\tau}_2)\prod_{i=1}^{k} P(a_i)^{\mu(a_i)}. \end{aligned}$$

The maximum likelihood of joint types $\overline{\overline{\tau}}$ is defined similarly to that of $(\overline{x}_1, \overline{x}_2)$ and the maximizing distributions are same if $\overline{\overline{\tau}} = \tau(\overline{x}_1, \overline{x}_2)$.

For closeness testing, it is useful to consider likelihood of $(\overline{x}_1, \overline{x}_2)$ under same pair of distributions $P_1 = P_2$. By a generalization and abuse of various notations, it is easy to see that if $P_3 = P_1 = P_2$,

$$P_{3,3}(\overline{x}_1, \overline{x}_2) = P_3(\overline{x}_1\overline{x}_2) = \prod_{i=1}^{k} P_3(a_i)^{\mu_1(a_i)+\mu_2(a_i)}$$

where the sequence $\overline{x}_1\overline{x}_2$ denotes the catenation of $\overline{x}_1$ and $\overline{x}_2$. We also conveniently define

$$\hat{P}_{3,3}(\overline{x}_1, \overline{x}_2) \overset{\text{def}}{=} \max_{P_3} P_{3,3}(\overline{x}_1, \overline{x}_2) = \prod_{i=1}^{k}\Big(\frac{\mu_1(a_i)+\mu_2(a_i)}{n_1+n_2}\Big)^{\mu_1(a_i)+\mu_2(a_i)}.$$

It follows that $\arg\max_{P_3} P_{3,3}(\overline{x}_1, \overline{x}_2) = \hat{P}_{\overline{x}_1\overline{x}_2} = \tau(\overline{x}_1\overline{x}_2)$.

2.3 Symmetric properties of distributions

The probability multiset of a distribution $P = (P(a_1), \ldots, P(a_k))$ on alphabet $\mathcal{A} = \{a_1, \ldots, k\}$ is the collection of its probability values

$$\mathcal{M}(P) \overset{\text{def}}{=} (p_1, \ldots, p_k) \overset{\text{def}}{=} \big\{P(a_1), \ldots, P(a_k)\big\},$$

where $p_1 \geq p_2 \geq \cdots \geq p_k$. Similarly, the probability multiset of two distributions $P_1 = (P_1(a_1), \ldots, P_1(a_k))$ and $P_2 = (P_2(a_1), \ldots, P_2(a_k))$ is the collection of pairs of probability values for the different symbols under the two distributions,

$$\mathcal{M}(P_1, P_2) \overset{\text{def}}{=} \big\{(p_{1,1}, p_{2,1}), \ldots, (p_{1,k}, p_{2,k})\big\} \overset{\text{def}}{=} \big\{(P_1(a_i), P_2(a_i)) : i = 1, 2, \ldots, k\big\}.$$

Note that $\mathcal{M}(P_1, P_2)$ is different from $(\mathcal{M}(P_1), \mathcal{M}(P_2))$ which does not convey the relationship between symbols of the two distributions (although is useful for testing symmetric properties of distribution multisets, like entropy). The multiset of $d \geq 3$ distributions $P_1, \ldots, P_d$ is defined similarly.

A property π of d distributions is a function that maps each $(P_1, P_2, \ldots, P_d)$ to a range of values. The range can be arbitrary – reals, integers, real vectors *etc.* with a distance measure $D(\cdot, \cdot)$ defined on the range. A property π is symmetric if it depends on $(P_1, \ldots, P_d)$ only through their probability multiset $\mathcal{M}(P_1, \ldots, P_d)$. For example, entropy $H(P) = \sum_{i=1}^{k} -P(a_i)\log(P(a_i))$ is a symmetric property of P and the L_1 distance $|P_1 - P_2| = \sum_{i=1}^{k} |P_1(a_i) - P_2(a_i)|$ is a symmetric property of (P_1, P_2). The probability multiset $\mathcal{M}(P) = (p_1, \ldots, p_k)$ in itself can be considered as an example of a vector valued symmetric property of P with D defined *e.g.,* as the sorted L_1 distance on distributions, $|P - P'|_1 \overset{\text{def}}{=} \sum_{i=1}^{\infty} |p_i - p'_i|$.

In general, the goal of a property testing problem, *i.e.,* testing a property π, is to find whether $\pi(P_1, \ldots, P_d) \in \mathcal{P}_1$ or $\mathcal{P}_2$ given $(\overline{X}_1, \ldots, \overline{X}_d) \sim P_1^{n_1} \times \cdots \times P_d^{n_d}$, where $\mathcal{P}_1$ and $\mathcal{P}_2$ are (disjoint) subsets of the range of π and given $\pi(P_1, \ldots, P_d) \in \mathcal{P}_1 \cup \mathcal{P}_2$. A test $\Delta = \Delta(\overline{X}_1, \ldots, \overline{X}_d)$ outputs 1 or 2 to indicate whether $\pi \in \mathcal{P}_1$ or $\pi \in \mathcal{P}_2$. Likewise, in a property estimation problem, given $(\overline{X}_1, \ldots, \overline{X}_d)$, we want to find an estimator $\phi = \phi(\overline{X}_1, \ldots, \overline{X}_d)$ of $\pi(P_1, \ldots, P_d)$ such such that $D(\phi, \pi(P_1, \ldots, P_d)) \leq \epsilon$. The error probability of a test Δ or estimator ϕ is the maximum over all $(P_1, \ldots, P_d)$ of the probability that the output is incorrect when $(\overline{X}_1, \ldots, \overline{X}_d) \sim (P_1, \ldots, P_d)$.

For testing and estimating symmetric properties from *i.i.d.* sequences, it is natural to look at the multiset of counts of various symbols under each distribution. This information is conveyed by profiles and are considered in the next section.

2.4 Patterns, profiles and their likelihoods

It is useful to consider patterns and profiles of sequences when they are generated *i.i.d.* and we are interested in *symmetric* or label-invariant properties of distributions that depend on the probability multiset of the distribution. Technically, it suffices to look at profiles when we consider *i.i.d.* sequences but patterns are useful for analyzing symmetric properties even under other classes of distributions *e.g.,* Markov, exchangeable. Patterns make the analysis and intuition of some of our schemes clearer. A combinatorial quantity that we call as pattern count of profiles comes in handy for approximating profile maximum likelihoods, which are difficult to compute in general. Using patterns, it easier to relate our methods to sequences, types, and schemes based on sequence maximum likelihood. Patterns are natural to consider for compressing sequences. They have been studied extensively in the context of universal compression of large alphabet sources in [49, 48], and most of the tests and estimators are motivated by the pattern maximum likelihood methods used there, suitably extended to tuples of sequences.

The pattern of a sequence $\overline{x}$ is defined as follows. For any sequence $\overline{z}$, let $\mathcal{A}(\overline{z})$ denote the set of symbols that appear in $\overline{z}$. The index $\imath_{\overline{x}}(a)$ of a symbol $a \in \mathcal{A}(\overline{x})$ is

$$\imath_{\overline{x}}(a) \overset{\text{def}}{=} \min\{|\mathcal{A}(x_1 x_2 \cdots x_i)| : 1 \le i \le n \text{ and } x_i = a\},$$

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

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

obtained by replacing the symbols in $\overline{x}$ by their respective indices, and thus in the order of their first appearances. Notice that pattern is a way of representing sequences without referring to the symbol identities. 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 (on all alphabets) is represented by Ψ^n. For example, $\Psi^1 = \{1\}$, $\Psi^2 = \{11, 12\}$ and $\Psi^3 = \{111, 112, 121, 122, 123\}$.

The *profile* of $\overline{x}$ conveys the multiset of counts of various symbols in $\overline{x}$ and equivalently conveys the number of symbols appearing a given number of times in it. We represent the profile as

$$\varphi(\overline{x}) \overset{\text{def}}{=} \overline{\varphi} \overset{\text{def}}{=} (\varphi_1, \varphi_2, \ldots, \varphi_n),$$

where

$$\varphi_\mu \overset{\text{def}}{=} \varphi_\mu(\overline{x}) \overset{\text{def}}{=} |\{a : \mu_{\overline{x}}(a) = \mu\}| = \sum_{i=1}^{k} \mathbb{1}_{[\mu(a_i)=\mu]}$$

is the prevalence of μ and is the number of symbols that appear μ times, for $\mu = 1, 2, \ldots, n$. An equivalent way of representing $\varphi(\overline{x})$ is using the collection of counts of symbols that have appeared in $\overline{x}$. Let $m \overset{\text{def}}{=} m(\overline{x}) \overset{\text{def}}{=} \sum_{i=1}^{n} \varphi_\mu$ be the number of symbols that appear in $\overline{x}$, and $\mu_1 \geq \mu_2 \geq \cdots \mu_m > 0$ be the counts of the symbols that appeared in $\overline{x}$. The multiplicity vector of $\overline{x}$ is

$$\mu(\overline{x}) \overset{\text{def}}{=} \overline{\mu} \overset{\text{def}}{=} \{\mu_1, \mu_2 \ldots, \mu_m\} \overset{\text{def}}{=} \{\mu(a) : \mu(a) > 0, a \in \mathcal{A}\}.$$

$\{\mu_1, \ldots, \mu_m\}$. (Curly braces are used to avoid confusion with the prevalence vector representation.)

Any valid profile of a length n sequence, $\overline{\varphi} = (\varphi_1, \ldots, \varphi_n)$ has a corresponding unique $\overline{\mu}$ in which the number of μ is φ_μ for $\mu = 1, 2, \ldots, n$. Likewise, any $\overline{\mu}$ has a corresponding $\overline{\varphi}$, and thus we use $\overline{\varphi}$ and its $\overline{\mu}$ with the same meaning. Furthermore, any such $\overline{\varphi}$ corresponds to an integer partition of n, since $\sum_{\mu=1}^{n} \mu\varphi_\mu = \sum_{i=1}^{m} \mu_i = n$. Thus, the set of all profiles of length-n sequences, denoted by Φ^n is in 1-1 correspondence with the (unordered) integer partitions of n. The following bound on $|\Phi^n|$ is due to a well known fact about partition number $p(n)$, the number of integer partitions of n, *e.g.,* see [31, 71, 49], and often useful for analyzing estimation properties of profile maximum likelihood.

Lemma 1. *For all $n > 2$,*

$$|\Phi^n| = p(n) \leq \exp\left(\pi\sqrt{\frac{2}{3}}\sqrt{n}\right). \qquad \square$$

The profile of a pattern $\overline{\psi} \in \Psi^n$ is defined similarly as that of any other sequence $\overline{x}$. The pattern count of a profile $\overline{\varphi}$ is the number of patterns which have the same profile $\overline{\varphi}$ and is denoted by

$$N(\overline{\varphi}) \overset{\text{def}}{=} |\{\overline{\varphi} : \varphi(\overline{\psi}) = \overline{\varphi}\}| = \frac{n!}{\prod_{\mu=1}^{n} (\mu!)^{\varphi_\mu} \varphi_\mu!}.$$

While the above combinatorial equality is easy to see by counting arguments [49], we show a more general result later on for *joint profiles* of multiple sequences.

As an example of some of the definitions above, the profile of $\mathtt{abdb}$ is $\varphi(\mathtt{abdb}) = (\varphi_1, \varphi_2, \varphi_3, \varphi_4) = (2, 1, 0, 0)$, indicating that there are 2 symbols ($\mathtt{a}, \mathtt{d}$) that appear once in $\mathtt{abdb}$ and 1 symbol ($\mathtt{b}$) that appears 2 times and 0 symbols that appear 3 and 4 times. In terms of multiset of counts, $\varphi(\mathtt{abdb}) = \{2, 1, 1\}$. The sequences $\mathtt{abdb}$ and $\mathtt{dcca}$ for example have the same profile, though their patterns are different.

Let P be a distribution on alphabet $\mathcal{A}$. The likelihood of a pattern $\overline{\psi} \in \Psi^n$ under P is the probability that a sequence $\overline{X} \sim P^n$ has pattern $\overline{\psi}$, given by

$$\begin{aligned} P(\overline{\psi}) &\stackrel{\text{def}}{=} P(\Psi(\overline{X}) = \overline{\psi}) = \sum_{\overline{x}:\Psi(\overline{x})=\overline{\psi}} P(\overline{x}) \\ &= \frac{1}{(k-m)!} \cdot \sum_{\sigma \in S_k} \prod_{i=1}^{k} p_{\sigma(i)}^{\mu_i} = \sum_{\sigma \in [k]^{\underline{m}}} \prod_{i=1}^{m} p_{\sigma(i)}^{\mu_i}, \end{aligned}$$

where $\mu(\overline{\psi}) = \{\mu_1, \mu_2, \ldots, \mu_m\}$ and $\mu_{m+1} = \cdots = \mu_k = 0$. Notice that $P(\overline{\psi})$ depends only on $\mathcal{M}(P)$ and $\varphi(\overline{\psi})$, and patterns with the same profile have the same probability under any given distribution. 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 the probability of the pattern 1213 is

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

The maximum likelihood of $\overline{\psi}$ is its maximum likelihood under all possible distributions on all alphabets (*i.e.,* over all $k \in \{1, 2, \ldots\}$),

$$\hat{P}(\overline{\psi}) \stackrel{\text{def}}{=} \max_{P} P(\overline{\psi}).$$

The maximizing distribution is denoted by $\hat{P}_{\overline{\psi}} \stackrel{\text{def}}{=} \arg\max_P P(\overline{\psi})$. We note that $\hat{P}_{\overline{\psi}}$ need not be its empirical distribution $\tau(\overline{\psi})$. For example, $P(112) = \frac{2}{9}$ when $P = (\frac{2}{3}, \frac{1}{3})$, whereas $\hat{P}(112) = \frac{1}{4}$ and $\hat{P}_{112} = (\frac{1}{2}, \frac{1}{2})$, *e.g.,* see [49].

Similarly, the likelihood of a profile $\overline{\varphi} \in \Phi^n$ under P is the probability that $\overline{X} \sim P^n$ has profile $\overline{\varphi}$, *i.e.,*

$$P(\overline{\varphi}) \stackrel{\text{def}}{=} P\big(\varphi(\overline{X}) = \overline{\varphi}\big) = \sum_{\overline{x}:\varphi(\overline{x})=\overline{\varphi}} P(\overline{x}).$$

As noted earlier, patterns with same profile $\overline{\varphi}$ have the same probability, hence

$$P(\overline{\varphi}) = N(\overline{\varphi})P(\overline{\psi}) = \frac{n!}{\prod_{\mu=1}^{n}(\mu!)^{\varphi_\mu}\varphi_\mu!}\sum_{\sigma\in[k]^{\underline{m}}}\prod_{i=1}^{m}p_{\sigma(i)}^{\mu_i},$$

where $\overline{\psi}$ is any pattern such that $\varphi(\overline{\psi}) = \overline{\varphi}$. We use $P(\Phi^n)$ to denote the distribution of $\varphi(\overline{X})$ when $\overline{X} \sim P^n$, *i.e.,* the distribution that assigns probability $P(\overline{\varphi})$ to each $\overline{\varphi} \in \Phi^n$.

The maximum likelihood of $\overline{\varphi}$ is

$$\hat{P}(\overline{\varphi}) \stackrel{\text{def}}{=} \hat{P}_{\overline{\varphi}}(\overline{\varphi}) \stackrel{\text{def}}{=} \max_P P(\overline{\varphi}),$$

its maximum likelihood under all possible distributions P on all alphabets, and the maximizing distribution is

$$\hat{P}_{\overline{\varphi}} \stackrel{\text{def}}{=} \arg\max_P P(\overline{\varphi}).$$

From the above discussion, it is clear that if a pattern $\overline{\psi}$ has profile $\overline{\varphi} = \varphi(\overline{\psi})$, then $\hat{P}_{\overline{\varphi}} = \hat{P}_{\overline{\psi}}$. So without loss of generality, we usually consider profile maximum likelihood since profiles are more natural to consider in the context of probability multiset estimation from *i.i.d.* sequences.

2.4.1 Joint patterns and profiles

Joint patterns

We extend the definition of patterns to two or more sequences for the purpose of closeness testing and in general, testing symmetric properties of several distributions. The *joint pattern*, or simply 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, for `bab` and `abca`, the first pattern is $\Psi(\texttt{bab}) = 121$ and that of the concatenated sequence is $\Psi(\texttt{bababca}) = 1212132$, hence the joint pattern is $\Psi(\texttt{bab}, \texttt{abca}) = (121, 2132)$. Clearly, the joint pattern conveys the patterns of the individual sequences and the association between the symbols of the sequences. The joint pattern of a tuple or 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, $\Psi^{2,1} = \{(11,1),(11,2),(12,1),(12,2),(12,3)\}$.

Joint profiles

The *joint profile*, or profile, of $\overline{x}_1, \overline{x}_2$ conveys the multiset of count pairs $(\mu_{\overline{x}_1}(a), \mu_{\overline{x}_2}(a))$ of various symbols that appear in $\overline{x}_1, \overline{x}_2$. Equivalently, it conveys the prevalences

$$\varphi_{\mu_1,\mu_2} \stackrel{\text{def}}{=} \varphi_{\mu_1,\mu_2}(\overline{x}_1, \overline{x}_2) \stackrel{\text{def}}{=} |\{a : \mu_{\overline{x}_1}(a) = \mu_1, \mu_{\overline{x}_2}(a) = \mu_2\}|,$$

i.e., the number of symbols that have appeared μ_1 times in $\overline{x}_1$ and μ_2 times in $\overline{x}_2$ for all $(\mu_1, \mu_2) \in \{0, 1, \ldots, n_1\} \times \{0, 1, \ldots, n_2\}$ and $\varphi_{0,0} \equiv 0$. We represent the profile of $(\overline{x}_1, \overline{x}_2)$ as

$$\varphi(\overline{x}_1, \overline{x}_2) \stackrel{\text{def}}{=} \overline{\overline{\varphi}} \stackrel{\text{def}}{=} \left[\varphi_{\mu_1,\mu_2}\right].$$

An equivalent way of representing the profile of $\overline{x}_1, \overline{x}_2$ is using the multiset of pairs of counts of each symbol that has appeared in the two sequences. The joint multiplicity vector of $\overline{x}_1, \overline{x}_2$ is

$$\begin{aligned} \mu(\overline{x}_1, \overline{x}_2) \stackrel{\text{def}}{=} \overline{\overline{\mu}} &\stackrel{\text{def}}{=} \left\{(\mu_{\overline{x}_1}(a), \mu_{\overline{x}_2}(a)) : a \in \mathcal{A} \text{ and } \mu_{\overline{x}_1}(a) > 0 \text{ or } \mu_{\overline{x}_2}(a) > 0\right\} \\ &\stackrel{\text{def}}{=} \{(\mu_{1,i}, \mu_{2,i}) : i = 1, 2, \ldots, m\}. \end{aligned}$$

Every joint profile $\overline{\overline{\varphi}}$ has a corresponding multiplicity vector $\overline{\overline{\varphi}}$ and vice versa. The profile of a joint pattern $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n_1,n_2}$ is defined similarly to that of any sequence pair. 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 prevalences φ_{μ_1,μ_2} are arranged in a matrix with the rows indexed with μ_1 and columns with μ_2. As we see above, $\varphi_{1,2} = 2$, since there are 2 symbols, `d` and `a`, that appear $\mu_1 = 1$ times in `dac` and $\mu_2 = 2$ times in `adbda`.

Number of profiles of a given length

We use $|\Phi^{n_1,n_2}|$ to denote the set of all profiles of sequence pairs of length (n_1, n_2). Similar to the case of $|\Phi^n|$, each profile $\overline{\overline{\varphi}} \in \Phi^{n_1,n_2}$ corresponds to a (unordered) partition of (n_1, n_2), *i.e.,*

$$\sum_{\mu_1=0}^{n_1}\sum_{\mu_2=0}^{n_2} \varphi_{\mu_1,\mu_2} \cdot (\mu_1, \mu_2) \;=\; \sum_{i=1}^{m} (\mu_{1,i}, \mu_{2,i}) \;=\; (n_1, n_2),$$

where the summation is performed componentwise. Similar to $|\Phi^n| \leq e^{3\sqrt{n}}$, we show that $|\Phi^{n_1,n_2}| \leq e^{3(n_1^{2/3}+n_2^{2/3})}$. We show the result in general for $|\Phi^{n_1,\ldots,n_d}| = p(n_1, \ldots, n_d)$, the number of different profiles of all d-tuples of sequences $(\overline{x}_1, \ldots, \overline{x}_d)$ of length $(n_1, \ldots, n_d)$. Here, $p(n_1, \ldots, n_d)$ is the number of integer partitions of $(n_1, \ldots, n_d)$, *i.e.,* the number of multisets of integer d-tuples $\{(\mu_{1,i}, \mu_{2,i}, \ldots, \mu_{d,i})\}_{i=1}^m$ such that $\sum_{i=1}^m \mu_{j,i} = n_j$ for $j = 1, 2, \ldots, d$. As an example, $p(2, 1) = 4$, since

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

Also see [21] for similar bounds on a related combinatorial structure called multidimensional partitions, where for each partition, the sum of all components of all parts is n, *i.e.,* $\sum_{i=1}^m \sum_{j=1}^d \mu_{i,j} = n$. We note that $|\Phi^{n_1,n_2}|$ does not factorize unlike the the number of joint types $|\mathcal{T}^{n_1,n_2}| = |\mathcal{T}^{n_1}||\mathcal{T}^{n_2}|$.

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

$$|\Phi^{n_1,\ldots,n_d}| \;=\; p(n_1, \ldots, n_d) \;\leq\; \exp\left(2\left(1 + \frac{1}{d}\right)\sum_{j=1}^{d} n_j^{d/(d+1)}\right).$$

Proof. The proof is similar to that for $p(n)$ in [71]. The (ordinary) generating function of $p(n_1, \ldots, n_d)$ is

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

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

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

In the Inequality (a), we consider the cases $l = 1$ and $l > 1$ separately. When $l > 1$, we have 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 $G(x_1, \ldots, x_d) > p(n_1, \ldots, n_d) x^{n_1} x^{n_2} \cdots x^{n_d}$,

$$\begin{aligned}
\log p(n_1, \ldots, n_d) &< \log G(x_1, \ldots, x_d) - \sum_{j=1}^{d} n_j \log x_j \\
&< \frac{2}{\prod_{j=1}^{d} (1 - x_j)} - \sum_{j=1}^{d} n_j \log x_j.
\end{aligned}$$

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

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

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 $\epsilon \leq \frac{1}{2}$, and therefore we have that $\log\big(1 - n_j^{-1/(d+1)}\big) \leq 2n_j^{-1/(d+1)}$ for $n_j > 2^{d+1}$ and $j = 1, \ldots, d$. □

Pattern counts of profiles

The pattern count of a joint profile $\overline{\overline{\varphi}} \in \Phi^{n_1,n_2}$ is the number of joint patterns which have the same profile $\overline{\overline{\varphi}}$ and is denoted by

$$N(\overline{\overline{\varphi}}) \stackrel{\text{def}}{=} |\{(\overline{\psi}_1, \overline{\psi}_2) : \varphi(\overline{\psi}_1, \overline{\psi}_2) = \overline{\overline{\varphi}}\}|.$$

For example, consider the profile $\overline{\overline{\varphi}} = \Phi(1232, 13)$ which has $\varphi_{1,1} = 2$, $\varphi_{2,0} = 1$ and all other $\varphi_{\mu_1,\mu_2} = 0$. Then, $N(\overline{\overline{\varphi}}) = 12$ since the set of all joint patterns that have this profile is $\{(1123, 23)$, $(1123, 32)$, $(1213, 23)$, $(1213, 32)$, $(1223, 13)$, $(1223, 31)$, $(1231, 23)$, $(1231, 32)$, $(1232, 13)$, $(1232, 31)$, $(1233, 13)$, $(1233, 21)\}$. Pattern counts of joint profiles, $N(\overline{\overline{\varphi}})$ do not factorize unlike sequence counts of joint types $N(\overline{\overline{\tau}}) = N(\overline{\tau}_1)N(\overline{\tau}_2)$. The next lemma shows an explicit combinatorial expression for $N(\overline{\overline{\varphi}})$ for the general case of $\overline{\overline{\varphi}} \in \Phi^{n_1,\ldots,n_d}$ as mentioned previously.

Lemma 3. *For all positive integers d and all $\overline{\overline{\varphi}} = [\varphi_{\mu_1,\ldots,\mu_d}] \in \Phi^{n_1,\ldots,n_d}$,*

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

Proof. We show the lemma for $d = 2$, and the proof is similar for any $d \geq 1$. Let $\overline{\overline{\varphi}} \in \Phi^{n_1,n_2}$. Any joint pattern $(\overline{\psi}_1, \overline{\psi}_2)$ that has profile $\overline{\overline{\varphi}}$ 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

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

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 $\overline{\overline{\varphi}}$ is

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

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

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

Clearly, $N^*(\overline{\overline{\varphi}}) = m! \cdot N(\varphi)$, 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 $\overline{\overline{\varphi}}$. Furthermore, the sets of sequence pairs generated in this way by different joint patterns are disjoint. So $N^*(\overline{\overline{\varphi}}) \geq m!{\cdot}N(\overline{\overline{\varphi}})$.

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

Thus,

$$N(\overline{\overline{\varphi}}) = \frac{N^*(\overline{\overline{\varphi}})}{m!} = \frac{n_1!n_2!}{\prod\limits_{\mu_1=0}^{n_1}\prod\limits_{\mu_2=0}^{n_2} (\mu_1!\mu_2!)^{\varphi_{\mu_1,\mu_2}}\varphi_{\mu_1,\mu_2}!}. \qquad \square$$

Likelihoods of joint profiles and patterns

Let (P_1, P_2) be a pair of distributions on $\mathcal{A}$, where $P_1 \overset{\text{def}}{=} (P_1(a_1), \ldots, P_1(a_k))$ and $P_2 \overset{\text{def}}{=} (P_2(a_1), \ldots, P_2(a_k))$. The probability of a joint pattern $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n_1,n_2}$ under (P_1, P_2) is the probability that $(\overline{X}_1, \overline{X}_2) \sim P_1^{n_1} \times P_2^{n_2}$ has joint pattern $(\overline{\psi}_1, \overline{\psi}_2)$, and is denoted by

$$\begin{aligned} P_{1,2}(\overline{\psi}_1, \overline{\psi}_2) &= P_{1,2}\big(\Psi(\overline{X}_1, \overline{X}_2) = (\overline{\psi}_1, \overline{\psi}_2)\big) = \sum_{\substack{(\overline{x}_1, \overline{x}_2): \\ \Psi(\overline{x}_1, \overline{x}_2) = (\overline{\psi}_1, \overline{\psi}_2)}} P_1(\overline{x}_1) P_2(\overline{x}_2) \\ &= \sum_{\sigma \in [k]^{\underline{m}}} \prod_{i=1}^{m} p_{1,\sigma(i)}^{\mu_{1,i}} p_{2,\sigma(i)}^{\mu_{2,i}}, \end{aligned}$$

where $\varphi(\overline{\psi}_1, \overline{\psi}_2) = \{(\mu_{1,i}, \mu_{2,i}) : i = 1, \ldots, m\}$. Clearly, $P_{1,2}(\overline{\psi}_1, \overline{\psi}_2)$ depends only on $\mathcal{M}(P_1, P_2)$ and $\varphi(\overline{\psi}_1, \overline{\psi}_2)$. For example, if $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$, $P_1 = (p_\mathtt{a}, p_\mathtt{b}, p_\mathtt{c}, p_\mathtt{d})$ and $P_2 = (p'_\mathtt{a}, p'_\mathtt{b}, p'_\mathtt{c}, p'_\mathtt{d})$, and $(\overline{\psi}_1, \overline{\psi}_2) = (12, 13)$, then

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

The probability of a joint profile $\overline{\overline{\varphi}} \in \Phi^{n_1,n_2}$ under (P_1, P_2) is

$$P_{1,2}(\overline{\overline{\varphi}}) \overset{\text{def}}{=} P_{1,2}\big(\varphi(\overline{X}_1, \overline{X}_2) = \overline{\overline{\varphi}}\big) = \sum_{\overline{x}_1, \overline{x}_2 : \varphi(\overline{x}_1, \overline{x}_2) = \overline{\overline{\varphi}}} P_1(\overline{x}_1) P_2(\overline{x}_2).$$

Joint patterns with same profile $\overline{\overline{\varphi}}$ have the same probability, hence

$$P_{1,2}(\overline{\overline{\varphi}}) = N(\overline{\overline{\varphi}}) P(\overline{\psi}_1, \overline{\psi}_2),$$

where $(\overline{\psi}_1, \overline{\psi}_2)$ is any joint pattern such that $\varphi(\overline{\psi}_1, \overline{\psi}_2) = \overline{\overline{\varphi}}$.

We use $P_{1,2}(\Phi^{n,n})$ to denote the distribution of $\varphi(\overline{X}_1, \overline{X}_2)$ when $\overline{X}_1, \overline{X}_2 \sim P_1^n \times P_2^n$, *i.e.,* the distribution that assigns probability $P_{1,2}(\overline{\overline{\varphi}})$ to each $\overline{\overline{\varphi}} \in \Phi^{n,n}$.

The probabilities of joint patterns and profiles do not factorize unlike the probabilities of sequence pairs $P_{1,2}(\overline{x}_1, \overline{x}_2) = P_1(\overline{x}_1) P_2(\overline{x}_2)$ and joint types $P_{1,2}(\overline{\overline{\tau}}) = P_1(\overline{\tau}_1) P_2(\overline{\tau}_2)$.

The maximum likelihood of a joint pattern $(\overline{\psi}_1, \overline{\psi}_2)$ is its maximum likelihood under all possible distributions on all alphabets (*i.e.,* over all $k \in \{1, 2, \ldots\}$),

$$\hat{P}(\overline{\psi}_1, \overline{\psi}_2) \overset{\text{def}}{=} \hat{P}_{1,2}(\overline{\psi}_1, \overline{\psi}_2) \overset{\text{def}}{=} \max_{P_1, P_2} P(\overline{\psi}_1, \overline{\psi}_2).$$

Likewise, the maximum likelihood of a joint profile $\overline{\overline{\varphi}}$ is

$$\hat{P}(\overline{\overline{\varphi}}) \stackrel{\text{def}}{=} \hat{P}_{1,2}(\overline{\overline{\varphi}}) \stackrel{\text{def}}{=} \max_{P_1,P_2} P(\overline{\overline{\varphi}}).$$

Clearly, if $\varphi(\overline{\psi}_2, \overline{\psi}_2)$, then $\hat{P}_{1,2}(\overline{\overline{\varphi}}) = N(\overline{\overline{\varphi}})\hat{P}_{1,2}(\overline{\psi}_2, \overline{\psi}_2)$ and both can be maximized by the same (P_1, P_2).

The following observation follows from the various definitions above and is relevant for closeness testing.

Observation 4. *If $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n_1,n_2}$, then $\overline{\psi}_1\overline{\psi}_2 \in \Psi^{n_1+n_2}$. If $P_3 = P_1 = P_2$, then (by an abuse of notation),*

$$P_{3,3}(\overline{\psi}_1, \overline{\psi}_2) = P_3(\overline{\psi}_1\overline{\psi}_2),$$

and hence

$$P_{3,3}(\varphi(\overline{\psi}_1, \overline{\psi}_2)) \;=\; \frac{N(\varphi(\overline{\psi}_1, \overline{\psi}_2))}{N(\varphi(\overline{\psi}_1\overline{\psi}_2))} P_3(\varphi(\overline{\psi}_1\overline{\psi}_2)).$$

In particular, for two sequences $(\overline{x}_1, \overline{x}_2) \in \mathcal{A}^n \times \mathcal{A}^n$, $P_{3,3}(\varphi(\overline{x}_1, \overline{x}_2)) \neq P_3(\varphi(\overline{x}_1\overline{x}_2))$ in general. □

2.5 Hypothesis testing

The following facts about hypothesis testing are well known, *e.g.,* see [44, 54]. Let P and Q be two distributions on a discrete alphabet $\mathcal{Z}$. In a simple hypothesis testing problem, a random variable Z is generated either $\sim P$ or $\sim Q$ with probability $(\frac{1}{2}, \frac{1}{2})$. A test Δ labels Z as either P or Q to indicate whether its generated by P or by Q respectively. The error probability of Δ is the probability that it labels Z incorrectly, *i.e.,*

$$P_{\text{e}}(\Delta) \stackrel{\text{def}}{=} P_{\text{e}}(\Delta, P, Q) \stackrel{\text{def}}{=} \frac{1}{2}P(\Delta(Z) = \mathsf{Q}) + \frac{1}{2}Q(\Delta(Z) = \mathsf{P}).$$

The L_1 distance between P and Q is

$$|P - Q| \stackrel{\text{def}}{=} \sum_{z\in\mathcal{Z}} |P(z) - Q(z)|$$

and the *testing affinity* [37] or affinity is

$$|P \wedge Q| \overset{\text{def}}{=} \sum_{z \in \mathcal{Z}} \min\{P(z), Q(z)\} = 1 - \frac{1}{2}|P - Q|.$$

Fact 5. *The test Δ^* given by $P(Z) \overset{\mathtt{P}}{\underset{\mathtt{Q}}{\gtrless}} Q(Z)$ has the minimum error probability and thus*

$$P_{\mathrm{e}}^*(P, Q) = \frac{1}{2} \sum_{z \in \mathcal{Z}} \min\{P(z), Q(z)\} = \frac{1}{2}|P \wedge Q| = \frac{1}{2} - \frac{1}{4}|P - Q|. \qquad \square$$

Equivalently, no test can distinguish P and Q with error probability $< \frac{1}{2}|P \wedge Q| = \frac{1}{2} - \frac{1}{4}|P - Q|$. Hence, in general, we say P and Q are distinguishable or indistinguishable to imply that $|P \wedge Q|$ is close to 0 or 1, or equivalently $|P - Q|$ is close to 2 or 0 respectively.

Instead of considering Z being generated from P or Q with equal probability or prior $(\frac{1}{2}, \frac{1}{2})$, we also consider the worst error probability under both cases,

$$\hat{P}_{\mathrm{e}}(\Delta, P, Q) \overset{\text{def}}{=} \max\{P(\Delta(Z) = \mathtt{Q}), Q(\Delta(Z) = \mathtt{P})\}.$$

We notice that

$$\hat{P}_{\mathrm{e}}^{*}(P, Q) = \min_{\Delta} \hat{P}_{\mathrm{e}}(\Delta)$$

is such that

$$\frac{1}{2}|P \wedge Q| \le \hat{P}_{\mathrm{e}}^{*}(P, Q) \le |P \wedge Q|.$$

Since we mostly consider small error probabilities (and the error probabilities diminish rapidly in sequence lengths), the tests we consider have similar error guarantees in both worst case and average case (equal prior). We define distinguishability of two distributions P and Q based on their $\hat{P}_{\mathrm{e}}^{*}$.

Definition 6. Two distributions P and Q are δ-distinguishable if $\hat{P}_{\mathrm{e}}^{*}(P, Q) \le \delta$, *i.e.*, if there is a test that can distinguish between them with worst error probability at most δ. In particular, P and Q are δ^*-distinguishable, where $\delta^* = \hat{P}_{\mathrm{e}}^{*}(P, Q)$. $\square$

In composite hypothesis testing problems, instead of two distributions P and Q, there are two classes of distributions $\mathcal{P}$ and $\mathcal{Q}$. Given Z that is generated randomly according to some $P \in \mathcal{P}$ or some $Q \in \mathcal{Q}$, we want to find which class it is. Notice that if P and Q are chosen according to some known priors on $\mathcal{P}$ or $\mathcal{Q}$, the problem reduces to simple hypothesis testing. In general, it is not easy to find the best test for such problems or even find accurate bounds on the least error probability. However, it is easy to see that the worst error probability of any test over $\mathcal{P}$ and $\mathcal{Q}$ is at least $\max_{P\in\mathcal{P},Q\in\mathcal{Q}} \hat{P_{\rm e}}^*(P,Q) \ge \frac{1}{2}\max_{P\in\mathcal{P},Q\in\mathcal{Q}} |P \wedge Q|$. Thus, a reasonable goal for these problems can be to find tests whose error probabilities are not much larger than this. This may not always be possible and the minimum error probability between $\mathcal{P}$ and $\mathcal{Q}$ can be much higher than this lower bound. Better lower bounds can be obtained by considering maximum over priors on $\mathcal{P}$ and/or $\mathcal{Q}$, *e.g.,* see [37, 56, 22], but are harder to analyze.

A commonly used test for composite hypothesis testing is the generalized likelihood ratio test (GLRT) $\hat{P}(Z) \underset{Q}{\overset{P}{\gtrless}} \hat{Q}(Z)$, where the maximum likelihoods under each of the classes, $\hat{P}(Z) \stackrel{\rm def}{=} \max_{P\in\mathcal{P}} P(Z)$ and $\hat{Q}(Z) \stackrel{\rm def}{=} \max_{Q\in\mathcal{Q}} Q(Z)$ are used as "plug-in" estimates of actual likelihoods, similar to the case of simple hypothesis testing. In Section 3.3, several variants and approximations of the GLRT are considered and shown to have an error probability of at most $\sqrt{\delta}\cdot|\mathcal{Z}|$, where $\delta = \max_{P\in\mathcal{P},Q\in\mathcal{Q}} \hat{P_{\rm e}}^*(P,Q)$, and this is often sufficient for our purposes.

2.6 Poissonization and tail bounds

It is evident from previous sections that we want to analyze the distributions of profiles of *i.i.d.* sequences. However, from Section 2.4, we observe that profile probabilities do not have a simple structure and are unwieldy symmetric polynomials in the probabilities of the distribution multiset. One widely used "trick" to help in the analysis to a large extent is to consider the following *Poisson* model. Let poi(λ) denote the Poisson distribution with mean λ and poi$(\lambda, i) \stackrel{\rm def}{=} \Pr(Z = i) \stackrel{\rm def}{=} \frac{\lambda^i e^{-\lambda}}{i!}$ where $Z \sim$ poi(λ). Unlike the usual *multinomial* model where the given *i.i.d.* sequences $\overline{X}$ are of length n, in the Poisson model, the length n' is distributed according to $\text{poi}(n)$. Thus, we use $\overline{X} \sim P^{\text{poi}(n)}$ to imply that we generate $n' \sim \text{poi}(n)$ and $\overline{X}$ is a sequence of n' independent samples distributed $\sim P$.

The key advantage in the Poisson model is the fact that the distributions of symbol counts are independent, *i.e.,* $\mu(a) \sim \text{poi}(nP(a))$ and independent of other $\mu(a')$, for $a, a' \in \mathcal{A}$. At the same time, finding good tests and estimators that have low error probability and use n samples is equivalent to finding good tests and estimators that use $\text{poi}(n)$ samples, which is due to the sharp concentration of $\text{poi}(\lambda)$ around its mean λ, and is implied in the following tail bounds [68, 69].

Observation 7. *(Also [68, Corollary 32].) For any* $\alpha \in (0.5, 1)$ *and sufficiently large* $\lambda > f(\alpha)$*, if* $X \sim \text{poi}(\lambda)$*,*

$$\Pr\big(|X-\lambda| \ge \lambda^{\alpha}\big) \;\le\; 2\exp\big(-\frac{3}{8}\lambda^{2\alpha-1}\big) \;\le\; \exp\big(-\lambda^{0.99(2\alpha-1)}\big).$$

For all $\epsilon \in (0,1]$ *and* $\lambda > \frac{1}{\epsilon^2(1-\epsilon)}$*, if* $X \sim \text{poi}(\lambda)$*,*

$$\Pr\big(|X-\lambda| \ge \epsilon\lambda\big) \;\le\; 2\exp(-\epsilon^2\lambda/3).$$

For $\alpha \ge 2$*, and sufficiently large* $\lambda \ge 2$*, if* $X \sim \text{poi}(\lambda)$*,*

$$\Pr\big(X \ge \alpha\lambda\big) \;\le\; \exp(-\alpha\lambda/6).$$

□

The well known Chernoff bounds that we use in the thesis are given below.

Fact 8. *(Chernoff bounds.) Let* $X = \sum_{i=1}^{n} Y_i$ *be a sum of independent 0,1 random variables* $Y_1, \ldots, Y_n$ *such that* $\Pr(Y_i = 1) = p_i$*. Let* $\mu = E[X] = \sum_i p_i$ *. Then,*

- *For* $\epsilon \in (0,1]$, $\Pr(X \le (1-\epsilon)\mu) \le \exp(-\mu\epsilon^2/2)$.
- *For* $\epsilon \in [0,1]$, $\Pr(X > (1+\epsilon)\mu) \le \exp(-\mu\epsilon^2/3)$ *and for* $\epsilon > 1$, $\Pr(X \ge (1+\epsilon)\mu) \le \exp(-\mu\epsilon/3)$. □

Chapter 3

Closeness Testing

The first problem of testing symmetric properties of distributions that we study in detail is that of testing closeness between two distributions. As before, let $\mathcal{A} \stackrel{\text{def}}{=} \{a_1, \dots, a_k\}$ be an alphabet of size k. And let $P_1 = (P_1(a_1), \dots, P_1(a_k))$ and $P_2 = (P_2(a_1), \dots, P_2(a_k))$ be two unknown distributions on $\mathcal{A}$. Given two length-n sequences $\overline{X}_1$ and $\overline{X}_2$ generated *i.i.d.* according to P_1 and P_2 respectively, we want to find whether P_1 and P_2 are same or different. A *closeness test* Δ labels the given sequences $(\overline{X}_1, \overline{X}_2)$ as either $\texttt{same}$ or $\texttt{diff}$ to indicate whether the distributions that generated them are believed to be same or different, *i.e.,* $\Delta : \mathcal{A}^n \times \mathcal{A}^n \to \{\texttt{same}, \texttt{diff}\}$. The error probability of Δ for any (P_1, P_2) is the probability that it labels a sequence pair they generate incorrectly, *i.e.,*

$$P_{\rm e}(\Delta, P_1, P_2) \stackrel{\text{def}}{=} \begin{cases} \Pr(\Delta(\overline{X}_1, \overline{X}_2) = \texttt{diff}) & \text{if } P_1, P_2 \text{ are same,} \\ \Pr(\Delta(\overline{X}_1, \overline{X}_2) = \texttt{same}) & \text{if } P_1, P_2 \text{ are different.} \end{cases}$$

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 characterize the performance of closeness tests, one defines two classes of pairs of distributions $\mathcal{P}_{\texttt{same}}$ and $\mathcal{P}_{\texttt{diff}}$ consisting of pairs of distributions that are considered to be same and different respectively. The error performance of a test Δ is then specified in terms of the maximum error probability over all $(P_1, P_2) \in \mathcal{P}_{\mathit{same}} \cup \mathcal{P}_{\mathit{diff}}$ and $(\overline{X}_1, \overline{X}_2) \sim P_1^n \times P_2^n$, *i.e.,*

$$P_{\mathrm{e}}(\Delta, \mathcal{P}_{\mathit{same}}, \mathcal{P}_{\mathit{diff}}) \stackrel{\text{def}}{=} \max_{(P_1,P_2) \in \mathcal{P}_{\mathit{same}}, \mathcal{P}_{\mathit{diff}}} P_{\mathrm{e}}(\Delta, P_1, P_2).$$

No guarantees are provided for other distributions, *i.e.,* $(P_1, P_2) \notin \mathcal{P}_{\mathit{same}} \cup \mathcal{P}_{\mathit{diff}}$.

A common way of parametrizing the performance of tests is to define $\mathcal{P}_{\mathit{same}}, \mathcal{P}_{\mathit{diff}}$ using a suitable distance $D(\cdot, \cdot)$ defined on distribution pairs, say $\mathcal{P}_{\mathit{same}} = \{(P_1, P_2) : P_1 = P_2\}$ and $\mathcal{P}_{\mathit{diff}} = \{(P_1, P_2) : D(P_1, P_2) \geq \epsilon\}$ for some $\epsilon > 0$, and then specify the sample complexity, *i.e.,* length of sequences n needed, in terms of the alphabet size k of the distributions being considered, to guarantee a small error probability, say $P_{\mathrm{e}}(\Delta, \mathcal{P}_{\mathit{same}}, \mathcal{P}_{\mathit{diff}}) \leq \delta \leq \frac{1}{4}$. Better tests thus require smaller $n = f(k)$. Note that if a test guarantees an error probability $\frac{1}{4}$ using n samples, then it it can be improved to any $\delta < \frac{1}{4}$ using $n' = \mathcal{O}(n \log(\frac{1}{\delta}))$ samples (by taking majority decision on the outputs of Δ on $\mathcal{O}(\log \frac{1}{\delta})$ instances of length-n sequence pairs). We can equivalently parametrize the performance by specifying the size of classes $\mathcal{P}_{\mathit{same}}$ and $\mathcal{P}_{\mathit{diff}}$, say in terms of k, on which Δ has low error probability, say $P_{\mathrm{e}}(\Delta, \mathcal{P}_{\mathit{same}}, \mathcal{P}_{\mathit{diff}}) \leq \delta \leq \frac{1}{4}$, using a pair of sequences of a given length n.

For example, Batu *et al.* in [8] provide a closeness test that can distinguish between all pairs of same distributions $\mathcal{P}_{\mathit{same}} = \{P_1 = P_2\}$ and those that are separated in L_1 distance, $\mathcal{P}_{\mathit{diff}} = \{|P_1 - P_2| \geq \epsilon\}$, when the alphabet size is at most k, using a pair of sequences of length $n = \mathcal{O}(k^{2/3} \cdot \epsilon^{-4} \cdot \log \frac{k}{\delta})$. Equivalently, their test guarantees low error probability δ using sequences of length n whenever $P_1 = P_2$ or $|P_1 - P_2| \geq \epsilon$ and $k = \mathcal{O}(n^{3/2} \cdot \epsilon^4 \cdot \log \frac{n}{\delta})$. Other recent results of a similar flavor by various researchers [70, 29, 35, 69] for large alphabet distributions that were mentioned in Section 1.1 and can be stated in the above manner.

A commonly used distance $D(\cdot, \cdot)$ is the L_1 distance between distributions $|P_1 - P_2| \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} |P_1(a) - P_2(a)|$. The is mainly because of the relationship of L_1 distance to other common distances like Hellinger, L_2, Jenson-Shannon Divergence, and strong implications for these distances. The L_1 distance is bigger than functions of most distances and thus, suitable for testing if two distributions are different. For example, if the alphabet size $k = \Omega(n)$, it is easy to find examples where $L_1 = \Omega(1)$ but $L_2 = o(1)$.

Before we go on to describe our choice of $\mathcal{P}_{same}$ and $\mathcal{P}_{diff}$, followed by our tests and their error guarantees, we consider the following simple well known closeness test that helps motivate the form of our tests.

A closeness test based on sequence maximum likelihood

It is easy to see that the closeness testing problem as described above is a composite hypothesis testing problem, briefly discussed in Section 2.5. We therefore consider a generalized likelihood ratio test on the given sequences. The likelihood ratio is

$$\frac{\max_{P_1=P_2} P(\overline{X}_1, \overline{X}_2)}{\max_{P_1,P_2} P(\overline{X}_1, \overline{X}_2)} = \frac{\hat{P}(\overline{X}_1\overline{X}_2)}{\hat{P}(\overline{X}_2)\hat{P}(\overline{X}_2)}.$$

This ratio is always less than 1 (since the domain of maximization is smaller in numerator compared to the denominator). However, it is easy to see using the arguments we show later, or otherwise, that when $P_1 = P_2$, this ratio is larger than $t = \frac{1}{n\binom{n+k-1}{k-1}^2}$, *i.e.,* not too small, with high probability $1 - o(1)$. Furthermore, if $|P_1 - P_2| \geq \epsilon$, this ratio is at most $2^{-n\epsilon^2/8} \ll t$ with probability $1 - o(1)$ if $k = o(n)$. Hence, the test Δ^{emp} given by

$$\frac{\hat{P}(\overline{X}_1\overline{X}_2)}{\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)} \underset{diff}{\overset{same}{\gtrless}} t$$

has low error probability whenever $P_1 = P_2$ or $|P_1 - P_2| \geq \epsilon$, and $k = o(n)$. Also see [35] for details of the arguments above.

But when $k = \Omega(n)$, empirical distribution may not be a good estimate of the underlying distribution to use a plug-in estimates for actual likelihoods in the likelihood ratio test. Thus, Δ^{emp} may not have low error probability, as shown in an example in [35] and in the following, simpler, example.

Example 9. For large n and $k = n^3$, let P_1, P_2 be such that $P_1(a_1) = 1$ and $P_1(a_2) = \cdots = P_1(a_k) = 0$, and $P_2(a_1) = 1/2$ and $P_2(a_2) = \cdots = P_2(a_k) = 1/(2(k-1))$. The two distributions are clearly very different and $|P_1 - P_2| = 1$. If $\overline{X}_1 \sim P_1^n$ and $\overline{X}_2 \sim P_2^n$, then consider two typical sequences $\overline{X}_1 = a_1^n$ and $\overline{X}_2 = a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1}$. In particular, by the Birthday problem, with high probability no symbol in $\{a_2, a_3, \ldots, a_k\}$ appears more than once in $\overline{X}_2$. It follows that

$$\frac{\hat{P}(\overline{X}_1\overline{X}_2)}{\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)} = \frac{\hat{P}(a_1^{\frac{3n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1})}{\hat{P}(a_1^n)\hat{P}(a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1})} = \frac{(\frac{3}{4})^{\frac{3n}{2}} (\frac{1}{2n})^{\frac{n}{2}}}{1^n \times (\frac{1}{2})^{\frac{n}{2}} (\frac{1}{n})^{\frac{n}{2}}} = \left(\frac{3}{4}\right)^{\frac{3n}{2}} \approx 0.65^n,$$

suggesting as it should that the sequences were generated by different distributions.

However, when both $\overline{X}_1$ and $\overline{X}_2$ are generated according to the same distribution, P_2, then a typical pair of sequences is $\overline{X}_1 = a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1}$ and $\overline{X}_2 = a_1^{\frac{n}{2}} a_{\frac{n}{2}+2} \cdots a_{n+1}$ where no symbol in $\{a_2, a_3, \ldots, a_k\}$ appears more than once in $\overline{X}_1\overline{X}_2$. Then,

$$\begin{aligned}\frac{\hat{P}(\overline{X}_1\overline{X}_2)}{\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)} &= \frac{\hat{P}(a_1^n a_2 a_3 \cdots a_{n+1})}{\hat{P}(a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1})\hat{P}(a_1^{\frac{n}{2}} a_{\frac{n}{2}+2} \cdots a_{n+1})} \\ &= \frac{(\frac{1}{2})^n (\frac{1}{2n})^n}{(\frac{1}{2})^{\frac{n}{2}} (\frac{1}{n})^{\frac{n}{2}} \times (\frac{1}{2})^{\frac{n}{2}} (\frac{1}{n})^{\frac{n}{2}}} = 2^{-n},\end{aligned}$$

an even lower ratio than when the distributions were different.

Thus, the GLRT test $\frac{\hat{P}(\overline{X}_1\overline{X}_2)}{\hat{P}(\overline{X}_1)\hat{P}(\overline{X}_2)} \overset{same}{\underset{diff}{\gtrless}} t$ cannot have low error probability for both (P_1, P_2) and (P_2, P_2) for any choice of the threshold t. Furthermore, note that if $\overline{X}_1, \overline{X}_2$ are both generated according to P_2, then $\overline{X}_1$, $\overline{X}_2$ have very different empirical distribution estimates than $\overline{X}_1\overline{X}_2$, breaking the intuition for small alphabets that their types or empirical distribution estimates should be similar, given that they are generated by the same distribution. □

3.1 A closeness test based on profile maximum likelihood

The empirical distribution and equivalently, maximum likelihood of sequences or types is a natural choice when we want to estimate the probabilities of specific symbols, *i.e.,* the complete distribution that includes the probability multiset and their mapping to the underlying alphabet. However, as we notice, the notion of closeness between (P_1, P_2) is a symmetric property of P_1, P_2 since it depends only on $\mathcal{M}(P_1, P_2)$. Thus, for estimating the (joint) probability multiset, it is natural to consider the joint pattern and profiles of sequences as our observations. We therefore consider the pattern- or profile-based likelihood ratio test $\Delta^{\hat{P}(\varphi)}$ defined as

$$\frac{\max_{P_1=P_2} P(\varphi(\overline{X}_1,\overline{X}_2))}{\max_{P_1,P_2} P(\varphi(\overline{X}_1,\overline{X}_2))} = \frac{\hat{P}_{3,3}(\varphi(\overline{X}_1,\overline{X}_2))}{\hat{P}_{1,2}(\varphi(\overline{X}_1,\overline{X}_2))} \underset{diff}{\overset{same}{\gtrless}} t,$$

where by an abuse of notation, $\hat{P}_{3,3}(\varphi(\overline{X}_1,\overline{X}_2)) \overset{\text{def}}{=} \max_{P_3=P_1=P_2} P_{1,2}(\varphi(\overline{X}_1,\overline{X}_2))$ and the threshold $t \le 1$ is a parameter and its choice will be revealed later on. The likelihood ratio can be also written in terms of pattern probabilities as

$$\frac{\hat{P}_{3,3}(\varphi(\overline{X}_1,\overline{X}_2))}{\hat{P}_{1,2}(\varphi(\overline{X}_1,\overline{X}_2))} = \frac{\hat{P}_{3,3}(\Psi(\overline{X}_1,\overline{X}_2))}{\hat{P}_{1,2}(\Psi(\overline{X}_1,\overline{X}_2))} = \frac{\hat{P}(\Psi(\overline{X}_1\overline{X}_2))}{\hat{P}(\Psi(\overline{X}_1,\overline{X}_2))}.$$

Similar to sequence-based GLRT, the main idea behind the profile-based GLRT is that when $P_1 = P_2$, the likelihood ratio is not too small and when P_1, P_2 are very different, the ratio is exponentially or near exponentially small. This is shown in Theorem 12 further along – the ratio is $\ge e^{-7n^{2/3}}$ with high probability when $P_1 = P_2$ and $\ll e^{-7n^{2/3}}$ when P_1, P_2 are very different. Revisiting Example 9, 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, $\hat{P}(\Psi(\overline{X}_1,\overline{X}_2)) = \hat{P}(1^n, 1^{\frac{n}{2}}23\cdots(\frac{n}{2}+1)) \ge 1 \cdot (\frac{1}{2})^{\frac{n}{2}}(\frac{1}{2})^{\frac{n}{2}} = (\frac{1}{2})^n$, 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 continuous alphabet or a large tail, similar to P_2. Also, using the result for PML of "skewed patterns" in [45], $\hat{P}(\Psi(\overline{X}_1\overline{X}_2)) = \hat{P}(1^{\frac{3n}{2}}23\cdots(\frac{n}{2}+1)) = (\frac{3}{4})^{\frac{3n}{2}}(\frac{1}{4})^{\frac{n}{2}}$, which is attained 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,

$$\frac{\hat{P}(\Psi(\overline{X}_1\overline{X}_2))}{\hat{P}(\Psi(\overline{X}_1,\overline{X}_2))} \le \frac{(\frac{3}{4})^{\frac{3n}{2}}(\frac{1}{4})^{\frac{n}{2}}}{(\frac{1}{2})^n} = \left(\frac{3}{4}\right)^{\frac{3n}{2}} < 0.65^n.$$

When $(\overline{X}_1,\overline{X}_2) \sim (P_2,P_2)$, consider the pair of typical sequences $(\overline{X}_1,\overline{X}_2) = (a_1^{\frac{n}{2}} a_2 a_3 \cdots a_{\frac{n}{2}+1}, a_1^{\frac{n}{2}} a_{\frac{n}{2}+2} \cdots a_{n+1})$. Again using the results in [45], we have that $\hat{P}(\Psi(\overline{X}_1,\overline{X}_2)) \le \hat{P}(\Psi(\overline{X}_1)\hat{P}(\Psi(\overline{X}_2)) = \hat{P}(1^{\frac{n}{2}}23\cdots(\frac{n}{2}+1))^2 = \left((\frac{1}{2})^{\frac{n}{2}}(\frac{1}{2})^{\frac{n}{2}}\right)^2 = (\frac{1}{2})^{2n}$, and $\hat{P}(\Psi(\overline{X}_1\overline{X}_2)) = \hat{P}(1^n23\cdots(n+1)) = (\frac{1}{2})^n(\frac{1}{2})^n = (\frac{1}{2})^{2n}$. Hence, in this case (along with the fact that this ratio is ≤ 1),

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

as we wanted before. Moreover, 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}.

To analyze the error of $\Delta^{\hat{P}(\varphi)}$, we define and motivate the following choice of $\mathcal{P}_{\mathtt{same}}$ and $\mathcal{P}_{\mathtt{diff}}$.

3.2 A distinguishability based distance criterion

Our choice of $\mathcal{P}_{\mathtt{same}}$ is the class that contains all same pairs of distributions, on all alphabet sizes $k \in \{1, 2, \ldots\}$, *i.e.,*

$$\mathcal{P}_{\mathtt{same}} \stackrel{\text{def}}{=} \{(P_3, P_3)\} \stackrel{\text{def}}{=} \{(P_1, P_2) : P_1 = P_2\}.$$

Clearly, this is a very natural choice. Our choice of $\mathcal{P}_{\mathtt{diff}}$ is motivated as follows.

3.2.1 Symmetric and profile-based tests

We argue that without loss of generality, we only need to consider tests that depend only on $(\overline{X}_1, \overline{X}_2)$ through its profile $\varphi(\overline{X}_1, \overline{X}_2)$. Similar arguments were used in [7],[9, Section 3.1.3]. Since closeness is a symmetric property of (P_1, P_2), we want tests that have low error probability for all (P_1, P_2) with the same multiset $\mathcal{M}(P_1, P_2)$, regardless of the specific way $\mathcal{M}(P_1, P_2)$ is mapped to $\mathcal{A}$ and the actual symbols we observe. Accordingly, 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.,*

$$P_{\text{e,sym}}(\Delta, P_1, P_2) \stackrel{\text{def}}{=} \max_{\sigma \in S_k} P_{\text{e}}(\Delta, P_1^{\sigma}, P_2^{\sigma}),$$

where for any $\sigma \in S_k$, $P_1^{\sigma}, P_2^{\sigma}$ are obtained from P_1, P_2 by the permutation σ of the alphabet so that $P_1^{\sigma}(a_i) = P_1(a_{\sigma(i)})$ and $P_2^{\sigma}(a_i) = P_2(a_{\sigma(i)})$ for $i = 1, \ldots, k$.

A *symmetric* closeness test is one 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 $\tilde{\Delta} : \Psi^{n,n} \to \{\mathtt{same}, \mathtt{diff}\}$. Hence, a symmetric test depends only the joint pattern of the sequences. Note that for a symmetric test Δ, $P_{\text{e,sym}}(\Delta, P_1, P_2) = P_{\text{e}}(\Delta, P_1, P_2)$

for all distribution pairs (P_1, P_2). The following observation shows that without loss of generality, we may limit ourselves to considering only symmetric closeness tests since they increase the error probability by a factor of at most 2 (which is a limitation of the fact that we consider only deterministic tests Δ).

Observation 10. *Let $\Delta : \mathcal{A}^n \times \mathcal{A}^n \to \{\mathtt{same}, \mathtt{diff}\}$ be any closeness test, possibly not symmetric. Then, there exists a symmetric test $\tilde{\Delta} : \mathcal{A}^n \times \mathcal{A}^n \to \{\mathtt{same}, \mathtt{diff}\}$ such that for all pairs of distributions (P_1, P_2) over $\mathcal{A}$,*

$$P^n_{\mathrm{e,sym}}(\tilde{\Delta}, P_1, P_2) \;\le\; 2 \cdot P_{\mathrm{e,sym}}(\Delta, P_1, P_2).$$

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.,* $\tilde{\Delta}(\overline{x}_1, \overline{x}_2) = \mathtt{majority}\{\Delta(\overline{x}'_1, \overline{x}'_2) : \Psi(\overline{x}'_1, \overline{x}'_2) = \Psi(\overline{x}_1, \overline{x}_2)\}$. Clearly, $P_{\mathrm{e}}(\tilde{\Delta}, P^{\sigma}_{1,2})$ is same for all permutations σ of $\mathcal{A}$. Thus, if P_1, P_2 are considered same,

$$\begin{aligned}
P_{\mathrm{e,sym}}(\tilde{\Delta}, P_1, P_2) &= P_{\mathrm{e}}(\tilde{\Delta}, P_1, P_2) \\
&= \frac{1}{k!} \sum_{\sigma \in S_k} P^n_{\mathrm{e}}(\tilde{\Delta}, P^{\sigma}_1, P^{\sigma}_2) \\
&= \frac{1}{k!} \sum_{\sigma \in S_k} \sum_{\substack{(\overline{x}_1, \overline{x}_2): \\ \tilde{\Delta}(\overline{x}_1, \overline{x}_2) = \mathtt{diff}}} P^{\sigma}_1(\overline{x}_1) P^{\sigma}_2(\overline{x}_2) \\
&= \sum_{\substack{(\overline{x}_1, \overline{x}_2): \\ \tilde{\Delta}(\overline{x}_1, \overline{x}_2) = \mathtt{diff}}} \frac{1}{k!} \sum_{\sigma \in S_k} P^{\sigma}_1(\overline{x}_1) P^{\sigma}_2(\overline{x}_2) \\
&\overset{(a)}{\le} 2 \sum_{\substack{(\overline{x}_1, \overline{x}_2): \\ \Delta(\overline{x}_1, \overline{x}_2) = \mathtt{diff}}} \frac{1}{k!} \sum_{\sigma \in S_k} P^{\sigma}_1(\overline{x}_1) P^{\sigma}_2(\overline{x}_2) \\
&= 2 \cdot \frac{1}{k!} \sum_{\sigma \in S_k} \sum_{\substack{(\overline{x}_1, \overline{x}_2): \\ \Delta(\overline{x}_1, \overline{x}_2) = \mathtt{diff}}} P^{\sigma}_1(\overline{x}_1) P^{\sigma}_2(\overline{x}_2) \\
&\le 2 \cdot \max_{\sigma \in S_k} \sum_{\substack{(\overline{x}_1, \overline{x}_2): \\ \Delta(\overline{x}_1, \overline{x}_2) = \mathtt{diff}}} P^{\sigma}_1(\overline{x}_1) P^{\sigma}_2(\overline{x}_2) \\
&= 2 \cdot P_{\mathrm{e,sym}}(\Delta, P_1, P_2),
\end{aligned}$$

where in (a), we note that all $(\overline{x}_1, \overline{x}_2)$ having the same joint pattern have the same probability $\frac{1}{k!} \sum_{\sigma \in S_k} P^{\sigma}_1(\overline{x}_1) P^{\sigma}_2(\overline{x}_2)$. A similar argument can be shown for the case when $P_1 \neq P_2$ are considered different. □

Similar to symmetric or pattern-based tests, we also define *profile-based tests* as those whose output depends only on the profile of the given sequences, *i.e.,* $\Delta(\overline{x}_1, \overline{x}_2) = \tilde{\Delta}(\varphi(\overline{x}_1, \overline{x}_2))$ for all $(\overline{x}_1, \overline{x}_2)$, where $\tilde{\tilde{\Delta}} : \Phi^{n,n} \to \{\mathtt{same}, \mathtt{diff}\}$. By a similar argument as that in Observation 10, (considering permutations $\sigma_1, \sigma_2 : [n] \to [n]$ of positions in the sequences instead of the alphabet symbols), we can consider only profile-based tests.

3.2.2 Distinguishable distribution pairs

Definition 11. Two distributions (P_1, P_2) are said to be (n, δ)-different for some n and $0 \leq \delta \leq 1$, if for all P_3, there exists a profile-based test $\Delta = \Delta(P_1, P_2, P_3)$ such that

$$P_{\mathrm{e}}(\Delta, P_1, P_2) \leq \delta \quad \text{and} \quad P_{\mathrm{e}}(\Delta, P_3, P_3) \leq \delta,$$

using sequence pairs of length n and while considering (P_1, P_2) as different and (P_3, P_3) as same. □

Following the discussion in Section 2.5, (P_1, P_2) is (n, δ)-different is same as saying $P_{1,2}(\Phi^{n,n})$ is δ-distinguishable from all $P_{3,3}(\Phi^{n,n})$. Notice that we allow different tests Δ for each P_1, P_2, P_3 to achieve an error probability of $\leq \delta$. We choose our $\mathcal{P}_{\mathtt{diff}}$ to contain all (n, δ)-different pairs (P_1, P_2), *i.e.,*

$$\mathcal{P}_{\mathtt{diff}} \stackrel{\text{def}}{=} \{(P_1, P_2) : (P_1, P_2) \text{ is } (n, \delta)\text{-different}\}.$$

Again, by the discussion in Section 2.5, $\mathcal{P}_{\mathtt{diff}}$ is the largest class of (P_1, P_2) for which we can hope for a profile-based test Δ to exist such that $P_{\mathrm{e}}(\Delta, \mathcal{P}_{\mathtt{same}}, \mathcal{P}_{\mathtt{diff}}) \leq \delta$. That is, if a pair (P_1, P_2) is not (n, δ)-different, there is some (P_3, P_3) such that for all profile-based tests Δ, either $P_{\mathrm{e}}(\Delta, P_3, P_3) > \delta$ or $P_{\mathrm{e}}(\Delta, P_1, P_2) > \delta$ and hence cannot be included in $\mathcal{P}_{\mathtt{diff}}$. While this $\mathcal{P}_{\mathtt{diff}}$ is the largest one could hope for, we expect $P_{\mathrm{e}}(\Delta, \mathcal{P}_{\mathtt{same}}, \mathcal{P}_{\mathtt{diff}})$ to be much larger than δ for any profile-based test Δ, since δ is the error bound when one is allowed to use different tests for different (P_1, P_2, P_3).

Our first main result is that the error probability of the profile-based GLRT $\Delta^{\hat{P}(\varphi)}$ is $P_{\mathrm{e}}(\Delta^{\hat{P}(\varphi)}, P_1, P_2) \le \sqrt{\delta} \cdot e^{6n^{2/3}}$, which is small when $\delta \le e^{-14n^{2/3}}$. Along with the above discussion, this implies that the $\Delta^{N(\varphi)}$ is competitive against any other test whose error probability is $\delta \le e^{-14n^{2/3}}$. That is, if there is a profile-based test Δ such that $P_{\mathrm{e}}(\Delta, P_1, P_2) \le e^{-14n^{2/3}}$ when $(P_1, P_2) \in \mathcal{P}_{\mathtt{same}} \cup \mathcal{P}'_{\mathtt{diff}}$, then $\mathcal{P}'_{\mathtt{diff}} \subset \mathcal{P}_{\mathtt{diff}}$ that consists of (n, δ)-different distributions where $\delta \le e^{-14n^{2/3}}$ and hence $P_{\mathrm{e}}(\Delta^{\hat{P}(\varphi)}, P_1, P_2) \le e^{-n^{2/3}}$ when $(P_1, P_2) \in \mathcal{P}_{\mathtt{same}} \cup \mathcal{P}_{\mathtt{diff}}$.

Theorem 12. *For all $n \ge 8$, all $0 < \delta \le \exp(-12n^{2/3})$, and all pairs distributions (P_1, P_2) that are either same or (n, δ)-different, the error probability of the test $\Delta^{\hat{P}(\varphi)}$ using threshold parameter $t = \sqrt{\delta}$,* i.e., $\frac{\hat{P}_{3,3}(\varphi(\overline{X}_1, \overline{X}_2))}{\hat{P}_{1,2}(\varphi(\overline{X}_1, \overline{X}_2))} \underset{\mathtt{diff}}{\overset{\mathtt{same}}{\gtrless}} \sqrt{\delta}$ *is*

$$P_{\mathrm{e}}(\Delta^{\hat{P}(\varphi)}, P_1, P_2) \le \sqrt{\delta} \cdot e^{6n^{2/3}}.$$

The theorem is proved in the next section by using a general result about the competitive properties of several tests that are variants of GLRT. Results of a similar flavor can be found in [23]. We also show a simpler test $\Delta^{\hat{P}_1(\varphi)}$ in the next section, that involves computing the PML of only a single pattern, $\hat{P}(\Psi(\overline{X}_1\overline{X}_2))$, and offers similar error guarantee as $\Delta^{\hat{P}(\varphi)}$.

3.3 Competitivity of GLRT for composite hypothesis testing

We recall the general composite hypothesis testing problem considered in Section 2.5. Let $\mathcal{Z}$ be a discrete alphabet of size $|\mathcal{Z}|$. Let $\mathcal{P}$ and $\mathcal{Q}$ be two collections of probability distributions, all of which are on alphabet $\mathcal{Z}$. Given a random variable Z distributed according a distribution that belongs to either $\mathcal{P}$ or $\mathcal{Q}$, we want to find out which of them it is, with minimum error. A test Δ outputs $\Delta(Z) = \mathtt{P}$ or $\mathtt{Q}$ to indicate the collection $\mathcal{P}$ or $\mathcal{Q}$ respectively, and is therefore a mapping $\Delta : \mathcal{Z} \to \{\mathtt{P}, \mathtt{Q}\}$. The error probability of Δ with respect to a distribution $P \in \mathcal{P}$ is

$$P_{\mathrm{e}}(\Delta, P) \stackrel{\text{def}}{=} P(\Delta(Z) = \mathtt{Q}) = \sum_{z \in \mathcal{Z} : \Delta(z) = \mathtt{Q}} P(z),$$

where $Z \sim P$. Similarly, for $Q \in \mathcal{Q}$,

$$P_{\mathrm{e}}(\Delta, Q) \overset{\text{def}}{=} Q(\Delta(Z) = \mathtt{P}),$$

where $Z \sim Q$.The error probability of Δ with respect to $\mathcal{P}$ and $\mathcal{Q}$ is its maximum error probability over all distributions in $\mathcal{P}$ and $\mathcal{Q}$,

$$P_{\mathrm{e}}(\Delta) \overset{\text{def}}{=} P_{\mathrm{e}}(\Delta, \mathcal{P}, \mathcal{Q}) \overset{\text{def}}{=} \max_{R \in \mathcal{P} \cup \mathcal{Q}} P_{\mathrm{e}}(\Delta, R) \;=\; \max\{\max_{P \in \mathcal{P}} P_{\mathrm{e}}(\Delta, P), \max_{Q \in \mathcal{Q}} P_{\mathrm{e}}(\Delta, Q)\}.$$

A commonly used test for this problem is the *generalized likelihood ratio test* (GLRT), also called simply the *likelihood ratio test*, which assigns Z to the class under which it has higher maximum likelihood. Let $\hat{P}(z) \overset{\text{def}}{=} \max_{P \in \mathcal{P}} P(z)$, $\hat{Q}(z) \overset{\text{def}}{=} \max_{Q \in \mathcal{Q}} Q(z)$ and $\hat{R}(z) \overset{\text{def}}{=} \max_{R \in \mathcal{R}} R(z)$ denote the maximum likelihood of each symbol $z \in \mathcal{Z}$ under $\mathcal{P}$, $\mathcal{Q}$, and $\mathcal{R} \overset{\text{def}}{=} \mathcal{P} \cup \mathcal{Q}$, respectively. Two versions of GLRT are commonly used in the literature, defined as follows.

Definition 13. The test GLRT-1 or G1 is given by

$$\Delta^{\mathrm{G1}}(a) \overset{\text{def}}{=} \begin{cases} \mathtt{P}, & \text{if } \hat{P}(z)/\hat{Q}(z) > 1 \\ \mathtt{Q}, & \text{if } \hat{P}(z)/\hat{Q}(z) \le 1, \end{cases}$$

for all $a \in \mathcal{A}$, and denoted in short as $\hat{P}(Z) \underset{\mathtt{Q}}{\overset{\mathtt{P}}{\gtrless}} \hat{Q}(Z)$. □

The test is motivated by the well known simple fact that for the problem of *simple hypothesis testing* where we want to find whether a sample X has been generated by a distribution P or a distribution Q (and both cases are equally likely), the test that minimizes the error probability is $P(Z) \underset{\mathtt{Q}}{\overset{\mathtt{P}}{\gtrless}} Q(Z)$. For composite hypothesis testing, since we do not know which distribution to consider from each class, it is natural to *plug-in* the maximum likelihood estimates in place of the actual likelihoods. Another commonly used test is the following one [44].

Definition 14 (GLRT-2)**.** The test GLRT-2 or G2 is given by

$$\Delta^{\mathrm{G2},t}(a) \overset{\text{def}}{=} \begin{cases} \mathtt{P}, & \text{if } \hat{P}(z)/\hat{R}(z) > t \\ \mathtt{Q}, & \text{if } \hat{P}(z)/\hat{R}(z) \le t, \end{cases}$$

for some real threshold $0 < t < 1$. □

The ratio $\hat{P}(z)/\hat{R}(z) \leq 1$, *i.e.,* $\hat{P}(z) \leq \hat{R}(z)$, since the maximization in $\hat{P}(a)$ is performed on a smaller set $\mathcal{P} \subset \mathcal{R}$. Hence, we need $t < 1$, otherwise $\Delta^{\text{G2},t}(z) \equiv \mathtt{Q}$ for all $z \in \mathcal{Z}$. A possible advantage of the test GLRT-2 is that it may be easier to compute $\hat{R}(z)$ as opposed to $\hat{Q}(z)$. As shown later in the analysis of the error probability of these tests, we may consider any $\mathcal{R}' \supset \mathcal{P} \cup \mathcal{Q}$ instead of $\mathcal{R}$. In particular, we can consider the degenerate case of $\mathcal{R}'$ being the set of all distributions on $\mathcal{Z}$ which leads to the test below.

Definition 15. The test GLRT-3 or G3 is

$$\Delta^{\text{G3},t}(z) \overset{\text{def}}{=} \begin{cases} \mathtt{P}, & \text{if } \hat{P}(z) > t \\ \mathtt{Q}, & \text{if } \hat{P}(z) \leq t, \end{cases}$$

for some $t < 1$. □

Sometimes, it may not be easy to find $\hat{P}(z)$ or $\hat{Q}(z)$ for all $a \in \mathcal{A}$, but we may have approximate estimates of them. We consider three such tests.

Definition 16. Let $f : \mathcal{Z} \to \mathbb{R}$ be such that for some reals $c_1 \leq c_2$,

$$c_1 \cdot \hat{P}(z) \ \leq f(z) \ \leq \ c_2 \cdot \hat{P}(z)$$

for all $z \in \mathcal{Z}$. The test GLRT-4 or G4 is

$$\Delta^{\text{G4},t}(z) \overset{\text{def}}{=} \begin{cases} \mathtt{P}, & \text{if } f(z) > t \\ \mathtt{Q}, & \text{if } f(z) \leq t, \end{cases}$$

for some $t < c_2$. □

Definition 17. Let $g : \mathcal{Z} \to \mathbb{R}$ be such that for some reals $c_3 \leq c_4$,

$$c_3 \cdot \hat{P}(z) \ \leq \ g(z) \ \leq \ c_4 \cdot \frac{\hat{P}(z)}{\hat{R}(z)}$$

for all $z \in \mathcal{Z}$. The test GLRT-5 or G5 is

$$\Delta^{\text{G5},t}(z) \overset{\text{def}}{=} \begin{cases} \mathtt{P}, & \text{if } g(z) > t \\ \mathtt{Q}, & \text{if } g(z) \leq t, \end{cases}$$

for some $t < c_4$. □

In GLRT-5, we can consider any $\mathcal{R}' \supset \mathcal{P} \cup \mathcal{Q}$ instead of $\mathcal{R}$. The extreme case of $\hat{R}(z) \equiv 1$ for all $z \in \mathcal{Z}$ results in GLRT-4.

Definition 18. Let $g : \mathcal{Z} \to \mathbb{R}$ be such that for some reals $c_5 \leq c_6$,

$$c_5 \cdot \frac{\hat{P}(z)}{\hat{Q}(z)} \leq h(z) \ \leq \ c_6 \cdot \frac{\hat{P}(z)}{\hat{Q}(z)}$$

for all $z \in \mathcal{Z}$. The test GLRT-6 or G6 is

$$\Delta^{\mathrm{G6},t}(z) \ \stackrel{\mathrm{def}}{=} \ \begin{cases} \mathtt{P}, & \text{if } h(z) > t \\ \mathtt{Q}, & \text{if } h(z) \leq t, \end{cases}$$

for some $t > 0$. □

The following lemma shows a competitive property about the error probability of the above six tests.

Lemma 19. *Let $\mathcal{P}$ and $\mathcal{Q}$ be such that for some $\delta \in [0,1]$, all $(P,Q) \in \mathcal{P} \times \mathcal{Q}$ are δ-distinguishable. In other words, for all $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$, supp there is a $\Delta = \Delta(P,Q)$ such that $P_{\mathrm{e}}(\Delta, P) \leq \delta$ and $P_{\mathrm{e}}(\Delta, Q) \leq \delta$. Then*

P1 $P_{\mathrm{e}}(\Delta^{G1}) \leq \delta \cdot |\mathcal{Z}|$.

P2 $P_{\mathrm{e}}(\Delta^{G2,\sqrt{\delta}}) \leq \sqrt{\delta} \cdot |\mathcal{Z}|$.

P3 $P_{\mathrm{e}}(\Delta^{G3,\delta}) \leq \delta \cdot |\mathcal{Z}|$.

P4 $P_{\mathrm{e}}(\Delta^{G4,c_2\delta}) \leq \frac{c_2}{c_1}\delta \cdot |\mathcal{Z}|$.

P5 $P_{\mathrm{e}}(\Delta^{G5,\sqrt{c_3 c_4 \delta}}) \leq \sqrt{\frac{c_4}{c_3}}\sqrt{\delta} \cdot |\mathcal{Z}|$.

P6 $P_{\mathrm{e}}(\Delta^{G6,\sqrt{c_5 c_6 \delta}}) \leq \sqrt{\frac{c_6}{c_5}} \cdot \delta \cdot |\mathcal{Z}|$.

Proof. The specific values of t for various tests can be derived by using the arguments below for general t and then choosing t to minimize error both when $Z \sim P \in \mathcal{P}$ or $Z \sim Q \in \mathcal{Q}$.

P1: For any $P \in \mathcal{P}$, when $Z \sim P$,

$$\begin{aligned} P_{\mathrm{e}}(\Delta^{\mathrm{G1}}, P) = &\ P\Big(\frac{\hat{P}(Z)}{\hat{Q}(Z)} \le 1\Big) \\ = &\ P\Big(\Big(\frac{\hat{P}(Z)}{\hat{Q}(Z)} \le 1\Big) \wedge (P(Z) > \delta)\Big) + P\Big(\Big(\frac{\hat{P}(Z)}{\hat{Q}(Z)} \le 1\Big) \wedge (P(Z) \le \delta)\Big) \\ \overset{\text{(a)}}{\le} &\ 0 + P\big(P(Z) \le \delta\big) \\ = &\ \sum_{z:P(z)\le\delta} P(z) \\ \le &\ |\mathcal{Z}| \cdot \delta \end{aligned}$$

where in (a), for the first term, we use that if $P(Z) > \delta$, since P is δ-distinguishable from all $Q \in \mathcal{Q}$, $Q(Z) \le \delta$ for all Q and this Z, and hence $\hat{Q}(Z) \le \delta$. Thus, $\frac{\hat{P}(Z)}{\hat{Q}(Z)} \ge \frac{P(Z)}{\hat{Q}(Z)} > \frac{\delta}{\delta} = 1$ and there is no error in this case.

For the case when $Z \sim Q \in \mathcal{Q}$, $P_{\mathrm{e}}(\Delta^{\mathrm{G1}}, Q) \le \delta \cdot |\mathcal{Z}|$ by a similar argument and symmetry.

P2: If $Z \sim P \in \mathcal{P}$, then

$$P_{\mathrm{e}}(\Delta^{\mathrm{G2},\sqrt{\delta}}, P) \;=\; P\Big(\frac{\hat{P}(Z)}{\hat{R}(Z)} \le \sqrt{\delta}\Big) \;\overset{\text{(a)}}{\le}\; P(P(Z) \le \sqrt{\delta}) \;\le\; |\mathcal{Z}| \cdot \sqrt{\delta},$$

where in (a), $\frac{\hat{P}(Z)}{\hat{R}(Z)} \le \sqrt{\delta} \Rightarrow P(Z) \le \hat{P}(Z) \le \sqrt{\delta} \cdot \hat{R}(Z) \le \sqrt{\delta}$.

If $Z \sim Q \in \mathcal{Q}$, then

$$P_{\mathrm{e}}(\Delta^{\mathrm{G2},\sqrt{\delta}}, Q) \;=\; Q\Big(\frac{\hat{P}(Z)}{\hat{R}(Z)} > \sqrt{\delta}\Big) \;\overset{\text{(a)}}{\le}\; Q(Q(Z) \le \sqrt{\delta}) \;\le\; |\mathcal{Z}| \cdot \sqrt{\delta},$$

where in (a), $\frac{\hat{P}(Z)}{\hat{R}(Z)} > \sqrt{\delta} \Rightarrow Q(Z) \le \sqrt{\delta}$. Otherwise, if $Q(Z) > \sqrt{\delta} \ge \delta$, then $\hat{P}(Z) > \sqrt{\delta} \cdot \hat{R}(Z) \ge \sqrt{\delta} \cdot Q(Z) > \delta$. Thus, there is a $P = \hat{P}_Z \in \mathcal{P}$ such that both $P(Z)$ and $Q(Z)$ are $> \delta$, contradicting that P, Q are δ-distinguishable.

P3: If $Z \sim P \in \mathcal{P}$, then

$$P_{\mathrm{e}}(\Delta^{\mathrm{G3},\delta}, P) \;=\; P(\hat{P}(Z) \le \delta) \;\le\; P(P(Z) \le \delta) \;\le\; |\mathcal{Z}| \cdot \delta.$$

If $Z \sim Q \in \mathcal{Q}$, then

$$P_{\mathrm{e}}(\Delta^{\mathrm{G3},\delta}, Q) \;=\; Q(\hat{P}(Z) > \delta) \;\overset{\text{(a)}}{\leq}\; Q(Q(Z) \leq \delta) \;\leq\; |\mathcal{Z}| \cdot \delta,$$

where in (a), $\hat{P}(Z) > \delta \Rightarrow Q(Z) \leq \delta$. Otherwise if $Q(Z) > \delta$, then $P = \hat{P}_Z \in \mathcal{P}$ and Q are not δ-distinguishable leading to a contradiction.

P4: This is similar to P3. If $Z \sim P \in \mathcal{P}$, then

$$P_{\mathrm{e}}(\Delta^{\mathrm{G4},c_2\delta}, P) \;=\; P(f(Z) \leq c_2\delta) \;\overset{\text{(a)}}{\leq}\; P\big(P(Z) \leq \frac{c_2}{c_1}\delta\big) \;\leq\; |\mathcal{Z}| \cdot \frac{c_2}{c_1}\delta,$$

since in (a), $f(Z) \leq c_2\delta \Rightarrow P(Z) \leq \hat{P}(Z) \leq \frac{f(Z)}{c_1} \leq \frac{c_2\delta}{c_1}$.

If $Z \sim Q \in \mathcal{Q}$, then

$$P_{\mathrm{e}}(\Delta^{\mathrm{G4},c_2\delta}, Q) \;=\; Q(f(Z) > c_2\delta) \;\overset{\text{(a)}}{\leq}\; Q\big(Q(Z) \leq \frac{c_2}{c_1}\delta\big) \;\leq\; |\mathcal{Z}| \cdot \frac{c_2}{c_1}\delta,$$

where in (a), $f(Z) > c_2\delta \Rightarrow Q(Z) \leq \frac{c_2}{c_1}\delta$. Otherwise, $Q(Z) > \frac{c_2}{c_1}\delta \geq \delta$ and $\hat{P}(Z) \geq \frac{f(Z)}{c_2} > \delta$, leading to a contradiction that $\hat{P}_Z$ and Q are not δ-distinguishable.

P5: This is similar to P2. If $Z \sim P \in \mathcal{P}$, then

$$P_{\mathrm{e}}(\Delta^{\mathrm{G5},\sqrt{c_3c_4\delta}}, P) \;=\; P\big(g(Z) \leq \sqrt{c_3c_4\delta}\big) \;\overset{\text{(a)}}{\leq}\; P\Big(P(Z) \leq \sqrt{\frac{c_4}{c_3}}\sqrt{\delta}\Big) \;\leq\; |\mathcal{Z}| \cdot \sqrt{\frac{c_3}{c_4}}\sqrt{\delta},$$

since in (a), $g(Z) \leq \sqrt{c_3c_4\delta} \Rightarrow P(Z) \leq \hat{P}(Z) \leq \frac{g(Z)}{c_3} \leq \sqrt{\frac{c_4}{c_3}\delta}$.

If $Z \sim Q \in \mathcal{Q}$, then

$$P_{\mathrm{e}}(\Delta^{\mathrm{G5},\sqrt{c_3c_4\delta}}, Q) \;=\; Q\big(g(Z) > \sqrt{c_3c_4\delta}\big) \;\overset{\text{(a)}}{\leq}\; Q\Big(Q(Z) \leq \sqrt{\frac{c_4}{c_3}}\sqrt{\delta}\Big) \;\leq\; |\mathcal{Z}| \cdot \sqrt{\frac{c_3}{c_4}}\sqrt{\delta},$$

where in (a), $g(Z) > \sqrt{c_3c_4\delta} \Rightarrow Q(Z) \leq \sqrt{\frac{c_4}{c_3}\delta}$. Otherwise, $Q(Z) > \sqrt{\frac{c_4}{c_3}\delta} \geq \delta$ and $\hat{P}(Z) \geq \frac{g(Z)\hat{R}(Z)}{c_4} \geq \frac{g(Z)Q(Z)}{c_4} > \frac{\sqrt{c_3c_4\delta}\sqrt{\delta}}{c_4} = \delta$, leading to a contradiction that $\hat{P}_Z$ and Q are not δ-distinguishable.

P6: This is similar to P1. When $Z \sim P \in \mathcal{P}$,

$$\begin{aligned}
P_{\mathrm{e}}(\Delta^{\mathrm{G6},\sqrt{c_5c_6}}, P) = \; & P\big(h(Z) \le \sqrt{c_5c_6}\big) \overset{\text{(a)}}{\le} P\Big(\frac{\hat{P}(Z)}{\hat{Q}(Z)} \le \sqrt{\frac{c_6}{c_5}}\Big) \\
= \; & P\Big(\Big(\frac{\hat{P}(Z)}{\hat{Q}(Z)} \le \sqrt{\frac{c_6}{c_5}}\Big) \wedge \Big(P(Z) > \sqrt{\frac{c_6}{c_5}}\delta\Big)\Big) \\
& \quad + P\Big(\Big(\frac{\hat{P}(Z)}{\hat{Q}(Z)} \le \sqrt{\frac{c_6}{c_5}}\Big) \wedge \Big(P(Z) \le \sqrt{\frac{c_6}{c_5}}\delta\Big)\Big) \\
\overset{\text{(b)}}{\le} \; & 0 + P\Big(P(Z) \le \sqrt{\frac{c_6}{c_5}}\delta\Big) \\
\le \; & |\mathcal{Z}| \cdot \sqrt{\frac{c_6}{c_5}}\delta
\end{aligned}$$

where in (a), we use that $c_5 \frac{\hat{P}(Z)}{\hat{Q}(Z)} \le h(Z)$. In (b), for the first term, we use that if $P(Z) > \sqrt{\frac{c_6}{c_5}}\delta > \delta$, then $\hat{Q}(Z) \le \delta$. Thus, $\frac{\hat{P}(Z)}{\hat{Q}(Z)} \ge \frac{P(Z)}{\hat{Q}(Z)} > \frac{\sqrt{\frac{c_6}{c_5}}\delta}{\delta} = \sqrt{\frac{c_6}{c_5}}$ and there is no error in this case.

The case when $Z \sim Q \in \mathcal{Q}$ is similar:

$$\begin{aligned}
P_{\mathrm{e}}(\Delta^{\mathrm{G6},\sqrt{c_5c_6}}, Q) = \; & P\big(h(Z) > \sqrt{c_5c_6}\big) \le Q\Big(\frac{\hat{P}(Z)}{\hat{Q}(Z)} > \sqrt{\frac{c_5}{c_6}}\Big) \\
= \; & Q\Big(\Big(\frac{\hat{P}(Z)}{\hat{Q}(Z)} > \sqrt{\frac{c_5}{c_6}}\Big) \wedge \Big(Q(Z) > \sqrt{\frac{c_6}{c_5}}\delta\Big)\Big) \\
& \quad + Q\Big(\Big(\frac{\hat{P}(Z)}{\hat{Q}(Z)} > \sqrt{\frac{c_5}{c_6}}\Big) \wedge \Big(Q(Z) \le \sqrt{\frac{c_6}{c_5}}\delta\Big)\Big) \\
\le \; & 0 + Q\Big(Q(Z) \le \sqrt{\frac{c_6}{c_5}}\delta\Big) \\
\le \; & |\mathcal{Z}| \cdot \sqrt{\frac{c_6}{c_5}}\delta.
\end{aligned}$$

□

Clearly, the above lemma also holds when there is a single test Δ such that $P_{\mathrm{e}}(\Delta, R) \le \delta$ for all $R \in \mathcal{P} \cup \mathcal{Q}$, since it is implied that all $(P, Q) \in \mathcal{P} \times \mathcal{Q}$ are δ-distinguishable.

Using the above general lemma, we prove Theorem 12.

Proof of Theorem 12. We observe that $\mathcal{P}_{\texttt{same}} = \{(P_3, P_3)\}$ and $\mathcal{P}_{\texttt{diff}} = \{(P_1, P_2) : (P_1, P_2) \text{ are } (n, \delta)\text{-different}\}$ satisfy the conditions of Lemma 19, *i.e.*, since we consider only profile-based tests, we consider $\Phi^{n,n}$ as $\mathcal{Z}$. Likewise, the distributions induced by $(P_3, P_3) \in \mathcal{P}_{\mathtt{same}}$ on $\Phi^{n,n}$ (*i.e.,* the distribution of $\varphi(\overline{X}_1, \overline{X}_2)$ where $(\overline{X}_1, \overline{X}_2) \sim P_3^n \times P_3^n$) can be considered as $\mathcal{P}$ and those induced by $(P_1, P_2) \in \mathcal{P}_{\mathtt{diff}}$ can be considered as $\mathcal{Q}$. Together with Lemma 2 that implies $|\mathcal{Z}| = |\Phi^{n,n}| \leq e^{6n^{2/3}}$, the theorem follows from P2 of Lemma 19. □

By using P3 of Lemma 19 for GLRT-3 of Definition 15 we see that the following simplified GLRT $\Delta^{\hat{P}_1(\varphi)}$ also has similar error guarantees as $\Delta^{\hat{P}(\varphi)}$.

Corollary 20. *For all $n \geq 8$, all $0 < \delta < \exp(-12n^{2/3})$, the test $\Delta^{\hat{P}_1(\varphi)}$ given by $\hat{P}_{3,3}(\varphi(\overline{X}_1, \overline{X}_2)) \underset{\mathtt{diff}}{\overset{\mathtt{same}}{\gtrless}} \delta$ has error probability $P_{\mathrm{e}}(\Delta^{\hat{P}_1(\varphi)}, \mathcal{P}_{\mathtt{same}}, \mathcal{P}_{\mathtt{diff}}) \leq \delta \cdot e^{6n^{2/3}}$.*

Proof Sketch. Similar to Theorem 12 and using P3 of Lemma 19. □

3.4 A closeness test based on pattern counts of profiles

We have noted earlier that computing maximum likelihood of patterns of even single sequences is difficult in general. We also saw this while applying $\Delta^{\hat{P}(\varphi)}$ to the simple cases in Example 9. By Observation 4, the maximum likelihood (ratio) in $\Delta^{\hat{P}_1(\varphi)}$ is

$$\hat{P}_{3,3}(\varphi(\overline{X}_1, \overline{X}_2)) = N(\Phi(\overline{X}_1, \overline{X}_2))\hat{P}(\Psi(\overline{X}_1\overline{X}_2)) = \frac{N(\Phi(\overline{X}_1, \overline{X}_2))}{N(\Phi(\overline{X}_1\overline{X}_2))}\hat{P}_{3,3}(\varphi(\overline{X}_1, \overline{X}_2)),$$

and involves the maximum likelihood of a pattern or profile of the single sequence $\overline{X}_1\overline{X}_2$, compared to computing the PML of both the joint pattern $\Psi(\overline{X}_1, \overline{X}_2)$ and of the single pattern $\Psi(\overline{X}_1\overline{X}_2)$ in $\Delta^{\hat{P}(\varphi)}$. Nevertheless, it still requires computing the PML of the single pattern $\Psi(\overline{X}_1\overline{X}_2)$, which may not be easy.

We therefore try to approximate the maximum likelihoods of patterns and profiles with the aim of using GLRT-5 of Definition 18 along with P5 of Lemma 19. In [49], it was shown that the probability estimator Q for $\overline{\psi} \in \Psi^n$, given by

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

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. Specifically, since

$$\hat{P}(\overline{\psi}) = \frac{\hat{P}(\varphi(\overline{\psi}))}{N(\varphi(\overline{\psi}))} \leq \frac{1}{N(\varphi(\overline{\psi}))} = |\Phi^n| \cdot Q(\overline{\psi}),$$

it follows that

$$Q(\overline{\psi}) \;\geq\; \hat{P}(\overline{\psi}) \cdot \frac{1}{|\Phi^n|} \;\geq\; \hat{P}(\overline{\psi}) \cdot \exp\Big(-\pi\sqrt{\frac{2}{3}}\sqrt{n}\Big).$$

By a similar argument, it follows that for all joint patterns $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n,n}$, the "inverse pattern count" estimator

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

is a good estimate of $\hat{P}(\overline{\psi}_1, \overline{\psi}_2)$, *i.e.*,

$$Q(\overline{\psi}_1, \overline{\psi}_2) \;\geq\; \hat{P}(\overline{\psi}_1, \overline{\psi}_2) \cdot e^{-6n^{2/3}}.$$

Thus, to approximate the likelihood ratio

$$L \stackrel{\text{def}}{=} L(\overline{\psi}_1, \overline{\psi}_2) \stackrel{\text{def}}{=} L(\varphi(\overline{\psi}_1, \overline{\psi}_2)) \stackrel{\text{def}}{=} \frac{\hat{P}(\overline{\psi}_1\overline{\psi}_2)}{\hat{P}(\overline{\psi}_1, \overline{\psi}_2)} = \frac{\hat{P}_{3,3}(\varphi(\overline{\psi}_1, \overline{\psi}_2))}{\hat{P}_{1,2}(\varphi(\overline{\psi}_1, \overline{\psi}_2))}$$

in $\Delta^{\hat{P}(\varphi)}$, where $(\overline{\psi}_1, \overline{\psi}_2) = \Psi(\overline{X}_1, \overline{X}_2)$, we use the combinatorial quantity

$$L' \stackrel{\text{def}}{=} L'(\overline{\psi}_1, \overline{\psi}_2) \stackrel{\text{def}}{=} L'(\varphi(\overline{\psi}_1, \overline{\psi}_2)) \stackrel{\text{def}}{=} \frac{N(\varphi(\overline{\psi}_1, \overline{\psi}_2))}{N(\varphi(\overline{\psi}_1\overline{\psi}_2))}$$

by replacing the maximum likelihood of patterns with the inverse pattern counts of their respective profiles. Note that $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n,n}$ and $\varphi(\overline{\psi}_1, \overline{\psi}_2) \in \Phi^{n,n}$ are a joint pattern and profile respectively, whereas $\overline{\psi}_1\overline{\psi}_2 \in \Psi^{2n}$ and $\varphi(\overline{\psi}_1\overline{\psi}_2) \in \Phi^{2n}$ are a single pattern and profile of length $2n$ respectively. Observe that $Q(\overline{\psi}_1, \overline{\psi}_2)$ and $Q(\overline{\psi}_1\overline{\psi}_2)$ are good upper bounds (within a subexponential factor) for $\hat{P}(\overline{\psi}_1, \overline{\psi}_2)$ and $\hat{P}(\overline{\psi}_1\overline{\psi}_2)$, so it is not immediate how L' approximates L. For using GLRT-5 of Definition 18, we require both an upper bound and lower bound on L'. We also keep in mind that to relate with the general composite hypothesis testing problem, we use the correspondence $\mathcal{Z} \leftrightarrow \Phi^{n,n}$. With these considerations, we first observe the following lower bound on L'.

Observation 21. *For all* $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n,n}$,

$$\hat{P}_{3,3}(\varphi(\overline{\psi}_1, \overline{\psi}_2)) \leq L'(\varphi(\overline{\psi}_1, \overline{\psi}_2)).$$

Proof. We have

$$\begin{aligned} L' &= \frac{N(\varphi(\overline{\psi}_1, \overline{\psi}_2))}{N(\varphi(\overline{\psi}_1 \overline{\psi}_2))} = \frac{N(\varphi(\overline{\psi}_1, \overline{\psi}_2))}{N(\varphi(\overline{\psi}_1 \overline{\psi}_2))} \frac{\hat{P}_{3,3}(\overline{\psi}_1, \overline{\psi}_2)}{\hat{P}_{3,3}(\overline{\psi}_1, \overline{\psi}_2)} \\ &= \frac{\hat{P}_{3,3}(\varphi(\overline{\psi}_1, \overline{\psi}_2))}{\hat{P}(\varphi(\overline{\psi}_1 \overline{\psi}_2))} \geq \hat{P}_{3,3}(\varphi(\overline{\psi}_1, \overline{\psi}_2)), \end{aligned}$$

where in the second last equality, we use $P_{3,3}(\overline{\psi}_1, \overline{\psi}_2) = P_3(\overline{\psi}_1 \overline{\psi}_2)$ from Observation 4, and in the last inequality, $\hat{P}(\varphi(\overline{\psi}_1 \overline{\psi}_2)) \leq 1$. □

Hence, it remains to show an upper bound on L' of the form suitable for using GLRT-5. We show the following upper bound on L' for using GLRT-5.

Lemma 22. *For all joint patterns* $(\overline{\psi}_1, \overline{\psi}_2) \in \Psi^{n,n}$,

$$\begin{aligned} L'(\varphi(\overline{\psi}_1, \overline{\psi}_2)) = \frac{N(\varphi(\overline{\psi}_1, \overline{\psi}_2))}{N(\varphi(\overline{\psi}_1 \overline{\psi}_2))} &\leq \frac{\hat{P}_{3,3}(\varphi(\overline{\psi}_1, \overline{\psi}_2))}{\hat{P}_{1,2}(\varphi(\overline{\psi}_1, \overline{\psi}_2))} \frac{(n!)^2 2^{2n}}{(2n)!} \\ &< \frac{\hat{P}_{3,3}(\varphi(\overline{\psi}_1, \overline{\psi}_2))}{\hat{P}_{1,2}(\varphi(\overline{\psi}_1, \overline{\psi}_2))} \cdot \sqrt{\pi n} e^{\frac{1}{6n}}. \end{aligned}$$

Before proceeding with the proof of above lemma, we immediately see that this lemma and Observation 21 imply that GLRT-5 of Definition 18 can be used along with P4 of Lemma 19 to obtain the following closeness test and show error guarantees similar to that for $\Delta^{\hat{P}(\varphi)}$ in Theorem 12 and $\Delta^{\hat{P}_1(\varphi)}$ in Corollary 20.

Theorem 23. *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, the closeness test* $\Delta^{N(\varphi)}$ *given by*

$$\frac{N(\varphi(\overline{X}_1, \overline{X}_2))}{N(\varphi(\overline{X}_1 \overline{X}_2))} \underset{\mathit{diff}}{\overset{\mathit{same}}{\gtrless}} \sqrt{\delta} \cdot \sqrt[4]{\pi n} e^{\frac{1}{12n}}$$

has error guarantee

$$P_{\mathrm{e}}(\Delta^{N(\varphi)}, P_1, P_2) < \sqrt{\delta} \cdot e^{6n^{2/3}} \sqrt{\pi n} e^{\frac{1}{6n}}.$$

Thus, $P_{\mathrm{e}}(\Delta^{N(\varphi)}, \mathcal{P}_{\mathit{same}}, \mathcal{P}_{\mathit{diff}}) < \sqrt{\delta} \cdot e^{6n^{2/3}} \sqrt{\pi n} e^{\frac{1}{6n}}$.

Proof. The proof is along the lines of Theorem 12 by correspondence with the general composite hypothesis testing problem. The test $\Delta^{N(\varphi)}$ can be seen to be of the form of GLRT-5 in Definition 18, with $g = L'$, $c_3 = 1$ by Observation 21 and $c_4 = \frac{(n!)^2 2^{2n}}{(2n)!} < \sqrt{\pi n} e^{\frac{1}{6n}}$ by Lemma 22. Thus, the result follows by using P4 of Lemma 19. □

The rest of this section is devoted to proving Lemma 22. We begin with a related result on the probabilities of joint types of *i.i.d.* sequence pairs. We use the notation introduced for types in Section 2.2. We define the *sum type* of a joint type $\overline{\overline{\tau}} = \big((\mu_1(a_i), \mu_2(a_i))\big)_{i=1}^k \in \mathcal{T}^{n,n}$ as $\tau_s(\overline{\overline{\tau}}) \stackrel{\text{def}}{=} \big(\mu(a_i)\big)_{i=1}^k \in \mathcal{T}^{2n}$, 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 $\overline{\tau} \in \mathcal{T}^{2n}$ under a pair of distributions (P_1, P_2) is the probability of the set of all types $\overline{\overline{\tau}} \in \mathcal{T}^{n,n}$ such that $\tau_s(\overline{\overline{\tau}}) = \overline{\tau}$, *i.e.*,

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

For any pair of distributions (P_1, P_2) over $\mathcal{K} \times \mathcal{K}$, $P_{1/2} \stackrel{\text{def}}{=} (P_1 + P_2)/2$ denotes the distribution over $\mathcal{K}$ such that $P_{1/2}(a_i) = (P_1(a_i) + P_2(a_i))/2$ for $i = 1, 2, \ldots, k$.

Observation 24. *For all types $\overline{\tau} \in \mathcal{T}^{2n}$ and all (p_1, p_2),*

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

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

$$\begin{aligned} P_{1,2}(\overline{\tau}) &= \sum_{\substack{\overline{\overline{\tau}} \in \mathcal{T}^{n,n}: \\ \tau_s(\overline{\overline{\tau}}) = \overline{\tau}}} P_{1,2}(\overline{\overline{\tau}}) \\ &= \sum_{\substack{(\mu_1(a_1), \ldots, \mu_1(a_k)): \\ 0 \leq \mu_1(a_i) \leq \mu(a_i) \text{ for } i=1,\ldots,k, \\ \text{and } \mu_1(a_1) + \cdots + \mu_1(a_k) = n}} n!n! \prod_{i=1}^k \frac{1}{\mu_1(a_i)!(\mu(a_i) - \mu_1(a_i))!} P_1(a_i)^{\mu_1(a_i)} P_2(a_i)^{\mu(a_i) - \mu_1(a_i)} \\ &= \frac{n!n!}{\prod_{i=1}^k \mu(a_i)!} \sum_{\substack{(\mu_1(a_1), \ldots, \mu_1(a_k)): \\ 0 \leq \mu_1(a_i) \leq \mu(a_i) \text{ for } i=1,\ldots,k, \\ \text{and } \mu_1(a_1) + \cdots + \mu_1(a_k) = n}} \prod_{i=1}^k \binom{\mu(a_i)}{\mu_1(a_i)} P_1(a_i)^{\mu_1(a_i)} P_2(a_i)^{\mu(a_i) - \mu_1(a_i)} \end{aligned}$$

$$\begin{aligned}
&\le \frac{n!n!}{\prod_{i=1}^k \mu(a_i)!} \sum_{\substack{(\mu_1(a_1),\ldots,\mu_1(a_k)):\\ 0\le\mu_1(a_i)\le\mu(a_i) \text{ for } i=1,\ldots,k}} \prod_{i=1}^k \binom{\mu(a_i)}{\mu_1(a_i)} P_1(a_i)^{\mu_1(a_i)} P_2(a_i)^{\mu(a_i)-\mu_1(a_i)}\\
&= \frac{n!n!}{\prod_{i=1}^k \mu(a_i)!} \prod_{i=1}^k \Big(\sum_{\mu_1(a_i)=0}^{\mu(a_i)} \binom{\mu(a_i)}{\mu_1(a_i)} P_1(a_i)^{\mu_1(a_i)} P_2(a_i)^{\mu(a_i)-\mu_1(a_i)} \Big)\\
&= \frac{n!n!}{\prod_{i=1}^k \mu(a_i)!} \prod_{i=1}^k (P_1(a_i)+P_2(a_i))^{\mu(a_i)}\\
&= \frac{(n!)^2 2^{2n}}{(2n)!} \binom{2n}{\mu(a_1),\mu(a_2),\ldots,\mu(a_k)} \prod_{i=1}^k \Big(\frac{P_1(a_i)+P_2(a_i)}{2}\Big)^{\mu(a_i)}\\
&= \frac{(n!)^2 2^{2n}}{(2n)!} P_{1/2}(\tau'). \qquad \square
\end{aligned}$$

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

Observation 25. *For all profiles $\overline{\varphi} \in \Phi^n$ and all distributions P,*

$$P(\overline{\varphi}) = \sum_{\overline{\tau}\in\mathcal{T}^n:\varphi(\overline{\tau})=\overline{\varphi}} P(\overline{\tau}).$$

Likewise, for all profiles $\overline{\overline{\varphi}} \in \Phi^{n_1,n_2}$ and all pairs of distributions (P_1,P_2),

$$P_{1,2}(\overline{\overline{\varphi}}) = \sum_{\overline{\overline{\tau}}\in\mathcal{T}^{n_1,n_2}:\varphi(\overline{\overline{\tau}})=\overline{\overline{\varphi}}} P_{1,2}(\overline{\overline{\tau}}). \qquad \square$$

The *sum profile* of a profile $\overline{\overline{\varphi}} \in \Phi^{n,n}$ is $\varphi_s(\overline{\overline{\varphi}}) \overset{\text{def}}{=} \varphi(\overline{x}_1\overline{x}_2) \in \Phi^{2n}$ where $(\overline{x}_1,\overline{x}_2)$ is any sequence pair whose profile is $\varphi(\overline{\psi}_1,\overline{\psi}_2) = \overline{\overline{\varphi}}$. Hence, if $\overline{\overline{\varphi}} = [\varphi_{\mu_1,\mu_2}]$, where $\mu_1 = 0,1,\ldots,n$ and $\mu_2 = 0,1,\ldots,n$, then $\varphi_s(\overline{\overline{\varphi}}) = (\varphi_1,\varphi_2,\ldots,\varphi_{2n})$ is given by $\varphi_\mu = \sum_{\mu_1+\mu_2=\mu} \varphi_{\mu_1,\mu_2}$. The probability of a (sum) profile $\overline{\varphi} \in \Phi^{2n}$ under a pair of distributions (P_1,P_2) is the probability $P_{1,2}$ assigns to the set of all profiles $\overline{\overline{\varphi}} \in \Phi^{n,n}$ such that $\varphi_s(\overline{\overline{\varphi}}) = \overline{\varphi}$, *i.e.*,

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

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

Lemma 26. *For all* $\overline{\varphi} \in \Phi^{2n}$ *and all* (P_1, P_2),

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

Proof. Using Observations 24 and 25,

$$\begin{aligned} P_{1,2}(\overline{\varphi}) &= \sum_{\substack{\overline{\overline{\varphi}} \in \Phi^{n,n}: \\ \varphi_s(\overline{\overline{\varphi}}) = \overline{\varphi}}} P_{1,2}(\overline{\overline{\varphi}}) \\ &= \sum_{\substack{\overline{\overline{\tau}} \in \mathcal{T}^{n,n}: \\ \varphi_s(\varphi(\overline{\overline{\tau}})) = \varphi(\tau_s(\overline{\overline{\tau}})) = \overline{\varphi}}} P_{1,2}(\overline{\overline{\tau}}) \\ &= \sum_{\substack{\overline{\tau} \in \mathcal{T}^{2n}: \\ \varphi(\overline{\tau}) = \overline{\varphi}}} P_{1,2}(\overline{\tau}) \\ &\leq \sum_{\substack{\overline{\tau} \in \mathcal{T}^{2n}: \\ \varphi(\overline{\tau}) = \overline{\varphi}}} \frac{(n!)^2 2^{2n}}{(2n)!} P_{1/2}(\overline{\tau}) \\ &= P_{1/2}(\overline{\varphi}) \frac{(n!)^2 2^{2n}}{(2n)!}. \end{aligned}$$

□

And we are ready to prove Lemma 22.

Proof of Lemma 22. Let (P_1, P_2) be a pair of distributions that maximizes $P_{1,2}(\overline{\psi}_1, \overline{\psi}_2)$, *i.e.*, $\hat{P}(\overline{\psi}_1, \overline{\psi}_2) = P_{1,2}(\overline{\psi}_1, \overline{\psi}_2)$. Note that $\varphi_s(\varphi(\overline{\psi}_1, \overline{\psi}_2)) = \varphi(\overline{\psi}_1 \overline{\psi}_2)$. Using Lemma 26, we have

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

Thus,

$$\frac{N(\varphi(\overline{\psi}_1,\overline{\psi}_2))}{N(\varphi(\overline{\psi}_1\overline{\psi}_2))} \le \frac{\hat{P}(\overline{\psi}_1\overline{\psi}_2)}{\hat{P}(\overline{\psi}_1\overline{\psi}_2)}\frac{(n!)^2 2^{2n}}{(2n)!} = \frac{\hat{P}_{3,3}(\varphi(\overline{\psi}_1,\overline{\psi}_2))}{\hat{P}_{1,2}(\varphi(\overline{\psi}_1,\overline{\psi}_2))}\frac{(n!)^2 2^{2n}}{(2n)!}. \qquad \square$$

3.5 Sample complexity implications

The error analysis results of Theorems 12 and 23 can be rephrased in terms of sample complexity. While Theorems 12 and 23 are applicable only when $\delta \le \exp(-14n^{2/3})$, this section partially addresses the general case when $\delta < \frac{1}{2}$.

Observation 27. *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*

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

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 $\le \delta$. 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 $\sum_{i=r+1}^{2r+1} \delta^i \binom{2r+1}{i}(1-\delta)^{2r+1-i} \le \delta^2 \cdot e^{-14((2r+1)n)^{2/3}}$. □

Corollary 28. *If (P_1, P_2) are (n, δ)-different for some $0 < \delta < \frac{1}{4}$, then they are also (n', δ')-different where $\delta' \le \delta^2 \cdot e^{-14n'^{2/3}}$ and $n' = \max\left\{19n, \frac{120000n^3}{(\log_2 \frac{1}{4\delta})^3}\right\}$. Furthermore if $\delta < e^{-19n^{2/3}}$, then $n' = 19n$.*

Hence, for such (P_1, P_2), the closeness test $\Delta^{N(\varphi)}$ guarantees an error probability $P_{\mathrm{e}}(\Delta^{N(\varphi)}, P_1, P_2) \le \delta$ using sequences of length n'.

Proof Sketch. Follows from Observation 27 and Theorem 23. □

3.6 Remarks

3.6.1 Sequences of unequal lengths

The results in this chapter are also valid when $(\overline{X}_1, \overline{X}_2)$ have unequal lengths n_1, n_2. The definition of (n, δ)-different distributions (P_1, P_2) can be extended to (n_1, n_2, δ)-different distributions by their distinguishability on Φ^{n_1,n_2}. The tests $\Delta^{\hat{P}(\varphi)}$, $\Delta^{\hat{P}_1(\varphi)}$ and $\Delta^{N(\varphi)}$ remain unchanged, with error guarantees now having the factor $e^{3(n_1^{2/3}+n_2^{2/3})}$ instead of $e^{6n^{2/3}}$ earlier. In the proofs of $\Delta^{N(\varphi)}$, *i.e.,* Lemma 22 and its sublemmas, instead of $P_{1/2} = (P_1 + P_2)/2$, one uses $P_\eta = \eta P_1 + (1-\eta)P_2 = \frac{n_1}{n_1+n_2}P_1 + \frac{n_2}{n_1+n_2}P_2$ where $\eta = \frac{n_1}{n_1+n_2}$ to make the Binomial theorem pass through in Lemma 24 and thus incurs a factor of $\frac{n_1!n_2!}{(n_1+n_2)!}\frac{(n_1+n_2)^{n_1+n_2}}{n_1^{n_1}n_2^{n_2}} = \mathcal{O}(\sqrt{\frac{n_1 n_2}{n_1+n_2}})$ instead of $\mathcal{O}(\sqrt{n})$ earlier.

3.6.2 Other problems of testing symmetric properties

It is easy to see that the PML-based GLRT for closeness testing can be generalized for other problems of testing symmetric properties, as given by the following lemma. We use the notation from Section 2.3.

Lemma 29. *Let π be a symmetric property of lists of d distributions $(P_1, \ldots, P_d)$. Let $\mathcal{R}$ be a subset of the range of π and let $\mathcal{P}_1 \overset{\text{def}}{=} \{(P_1, \ldots, P_d) : \pi(P_1, \ldots, P_d) \in \mathcal{R}\}$. A list of distributions $(P'_1, \ldots, P'_d)$ is $(n_1, \ldots, n_d, \delta)$-different if $P'_{1,\ldots,d}(\Phi^{n_1,\ldots,n_d})$ is δ-distinguishable from all $P_{1,\ldots,d}(\Phi^{n_1,\ldots,n_d})$ where $(P_1, \ldots, P_d) \in \mathcal{P}_1$,* i.e., *for each $(P_1, \ldots, P_d) \in \mathcal{P}_1$, there is a test that can distinguish between $(P'_1, \ldots, P'_d)$ and $(P_1, \ldots, P_d)$ when given $\varphi(\overline{X}_1, \ldots, \overline{X}_d)$ where $(\overline{X}_1, \ldots, \overline{X}_d) \sim (P_1^{n_1}, \ldots, P_d^{n})$ or $\sim (P'^{n_1}_1, \ldots, P'^{n_d}_d)$ with error probability δ. Let $\mathcal{P}_2$ consists of all $(n_1, \ldots, n_d, \delta)$-different lists of distributions.*

Then, given $(\overline{X}_1, \ldots, \overline{X}_d) \sim (P_1, \ldots, P_d) \in \mathcal{P}_1 \cup \mathcal{P}_2$, the PML-based GLRT $\max_{(P_1,\ldots,P_d)\in\mathcal{P}_1} P_{1,\ldots,d}(\varphi(\overline{X}_1, \ldots, \overline{X}_d)) \underset{2}{\overset{1}{\gtrless}} \delta$ *has an error probability that is at most* $\delta \cdot |\Phi^{n_1,\ldots,n_d}| \leq \delta \cdot e^{2(1+\frac{1}{d})\sum_{j=1}^{d} n_j^{\frac{d}{d+1}}}$.

In terms of sample complexity, when $n_1 = n_2 = \cdots = n_d = n$, i.e., *$\mathcal{P}_2$ consists of $(n, \ldots, n, \delta)$-different distributions and $\delta < \frac{1}{4}$, the PML-based GLRT has* *error probability at most δ using sequences of length $n' = \mathcal{O}(\max\{n, \frac{n^{d+1}}{\log^{d+1}(\frac{1}{4\delta})}\})$.*

Proof Sketch. Follows from P3 of Lemma 19 and Lemma 2. The sample complexity argument is similar to that for Observation 27 and Corollary 28. □

The above lemma and the PML-based GLRT are therefore useful whenever the maximum likelihood in the GLRT, $\max_{(P_1,\ldots,P_d)\in\mathcal{P}_1} P_{1,\ldots,d}(\varphi(\overline{X}_1,\ldots,\overline{X}_d))$ can be computed or approximated efficiently. In several problems, both these conditions are satisfied. For example, consider the following corollary for testing uniformity of distributions.

Corollary 30. *Let $U[k]$ denote the uniform distribution on k symbols. Let $\mathcal{P}_1 = \{U[k] : k = 1, 2, \ldots,\}$ be the set of all uniform distributions. Suppose we are given $\overline{X} \sim P^n$ where P is either a uniform distribution or a distribution that is very different from uniform so that it can be distinguished from any uniform distribution with error probability at most $\delta \le e^{-4n^{1/2}}$. The test*

$$\max_{P\in\mathcal{P}_1} P(\varphi(\overline{X})) = N(\varphi(\overline{X}))\max_k \frac{k^{\underline{m(\overline{X})}}}{k^n} \underset{2}{\overset{1}{\gtrless}} \delta$$

(where output 1 indicates uniform and 2 indicates non-uniform) has error probability at most $\delta \cdot e^{3n^{1/2}} \le e^{-n^{1/2}}$. □

Note that the quantity $f(n,m) \overset{\text{def}}{=} \max_k(k^{\underline{m}}/k^n)$ in the above corollary can be computed easily. In other property testing problems, computing the PML under $\mathcal{P}_1$ in Lemma 29 may not be easy, *e.g.,* for testing whether entropy $H(P)$ is $< \alpha$ or $> \beta$ given alphabet size bounds.

Acknowledgement

Chapter 3 is adapted from Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, "Competitive closeness testing", *Journal of Machine Learning Research - Proceedings Track (COLT 2011)*, 2011. The dissertation author was the primary researcher and author of this paper.

Chapter 4

Classification

In this chapter, we consider the problem of classification which is related to closeness testing and has a very wide range of applications. The goal here is to classify *test* data into one among several classes characterized by *training* data belonging to them. For simplicity, we consider here the case of binary classification, where there are two classes, although most results extend to multiple classes as well. Let P_1 and P_2 be two unknown distributions on alphabet $\mathcal{A}$. We are given two length-n training sequences $\overline{X}_1 \sim P_1^n$ and $\overline{X}_2 \sim P_2^n$. Given a third test sequence $\overline{Y}$ of length n, distributed either $\sim P_1^n$ or $\sim P_2^n$, we want to find which one of them generated it. A classifier $\Gamma : \mathcal{A}^n \times \mathcal{A}^n \times \mathcal{A}^n \to \{1, 2\}$ outputs either $\Gamma(\overline{X}_1, \overline{X}_2, \overline{Y}) = 1$ or 2 to indicate whether $\overline{Y}$ is generated by P_1 or P_2. The error probability of the classifier is the probability that it classifies $\overline{Y}$ incorrectly. We consider the worst error probability over both cases $\overline{Y} \sim P_1^n$ and $\overline{Y} \sim P_2^n$, *i.e.,*

$$P_{\mathrm{e}}(\Gamma, P_1, P_2) \stackrel{\mathrm{def}}{=} \max\Big\{P_{1,2,1}\big(\Gamma(\overline{X}_1, \overline{X}_2, \overline{Y}) = 2\big),\ P_{1,2,2}\big(\Gamma(\overline{X}_1, \overline{X}_2, \overline{Y}) = 1\big)\Big\},$$

where, by an abuse of notation as in earlier chapters, $P_{1,2,1}$ denotes the distribution $(\overline{X}_1, \overline{X}_2, \overline{Y}) \sim P_1^n \times P_2^n \times P_1^n$ and $P_{1,2,2}$ is defined similarly.

Since it is natural to require that classifiers have the same error performance for all (P_1, P_2) that have the same multiset $\mathcal{M}(P_1, P_2)$, regardless of the specific way they $\mathcal{M}(P_1, P_2)$ is associated with the alphabet, we only consider symmetric classifiers that depend on $\overline{X}_1, \overline{X}_2, \overline{Y}$ only through their joint pattern $\Psi(\overline{X}_1, \overline{X}_2, \overline{Y})$. Furthermore, since sequences are generated *i.i.d.*, we only consider profile-based classifiers that depend on the given sequences through their joint profile $\varphi(\overline{X}_1, \overline{X}_2, \overline{Y})$. This follows from arguments similar to that in Subsection 3.2.1 or that in [9, Section 3.1.3].

Following the notation in Section 2.5, classification can therefore be considered as a composite hypothesis testing problem on the alphabet $\mathcal{Z} = \Phi^{n,n,n} = \{\varphi(\overline{X}_1, \overline{X}_2, \overline{Y})\}$ where the classes of distributions on $\Phi^{n,n,n}$ are

$$\mathcal{P}_1 \overset{\text{def}}{=} \{P_{1,2,1}(\Phi^{n,n,n})\} \text{ and } \mathcal{P}_2 \overset{\text{def}}{=} \{P_{1,2,2}(\Phi^{n,n,n})\},$$

i.e., the distributions induced on $\Phi^{n,n,n}$ by $\varphi(\overline{X}_1, \overline{X}_2, \overline{Y})$ when $(\overline{X}_1, \overline{X}_2, \overline{Y}) \sim P_1^n \times P_2^n \times P_1^n$ and $(\overline{X}_1, \overline{X}_2, \overline{Y}) \sim P_1^n \times P_2^n \times P_2^n$ respectively, by various P_1, P_2. Following the discussion in Section 2.5 and similar to the definition of (n, δ)-different distribution pairs considered in Subsection 3.2.2, we define (n, δ)-classifiable distributions as follows.

Definition 31. Two distributions P_1 and P_2 are said to be (n, δ)-classifiable if $P_{1,2,1}(\Phi^{n,n,n})$ is δ-distinguishable from all $P_{3,4,4}(\Phi^{n,n,n})$ and $P_{1,2,2}(\Phi^{n,n,n})$ is δ-distinguishable from $P_{3,4,3}(\Phi^{n,n,n})$.

In other words, for all P_3, P_4 there is a profile-based test that can distinguish between $\varphi(\overline{X}_1, \overline{X}_2, \overline{Y})$ where $(\overline{X}_1, \overline{X}_2, \overline{Y}) \sim P_1^n \times P_2^n \times P_2^n$ or $\sim P_3^n \times P_4^n \times P_4^n$ with error probability at most δ. Similarly, for all P_3, P_4 there is a profile-based test that can distinguish between $\varphi(\overline{X}_1, \overline{X}_2, \overline{Y})$ where $(\overline{X}_1, \overline{X}_2, \overline{Y}) \sim P_1^n \times P_2^n \times P_2^n$ or $\sim P_3^n \times P_4^n \times P_2^n$ with error probability at most δ. □

The motivation is the same as in Section 2.5 and Subsection 3.2.2. If two distributions P_1, P_2 are not (n, δ)-classifiable, then for some P_3, P_4, say $P_{1,2,1}(\Phi^{n,n,n})$ is not δ-distinguishable from $P_{3,4,4}(\Phi^{n,n,n})$. Hence, when given $(\overline{X}_1, \overline{X}_2, \overline{Y}) \sim P_{1,2,1}$ one cannot say reliably with error probability $\leq \delta$ whether $\overline{Y}$ was generated by the same distribution as $\overline{X}_1$, *e.g.,* by $P_{1,2,1}$ or by the same distribution as $\overline{X}_2$, *e.g.,* by $P_{3,4,4}$. Thus, (n, δ)-classifiable distribution pairs are the only pairs for which one can hope to have classification error probability of at most δ. And we want to find classifiers whose error probability is not much larger than δ for (n, δ)-classifiable distribution pairs.

4.1 Closeness testing and classification

It is easy to see the relationship between classification and closeness testing. One can simply test whether $\overline{X}_1, \overline{Y}$ are generated by the same distribution or not using a closeness test and output 1 or 2 respectively. We therefore have the following observations.

Observation 32. *If two distributions (P_1, P_2) are (n, δ)-different, they are also (n, δ)-classifiable.*

Proof Sketch. Clearly $P_{1,2,1}(\Phi^{n,n,n})$ is δ-distinguishable from all $P_{3,4,4}(\Phi^{n,n,n})$, since they can be distinguished using the second and third components of their profiles with error probability at most δ which is in turn because $P_{2,1}(\Phi^{n,n})$ is δ-distinguishable from $P_{4,4}(\Phi^{n,n})$ due to the fact that (P_1, P_2) are (n, δ)-different. Similarly, $P_{1,2,2}(\Phi^{n,n,n})$ is δ-distinguishable from all $P_{3,4,3}(\Phi^{n,n,n})$ using the first and third components of $\overline{\overline{\varphi}} \in \Phi^{n,n,n}$. □

Observation 33. *For all distribution pairs P_1, P_2 that are (n, δ)-different, the classifier $\Gamma^{\Delta^{N(\varphi)}}$ based on the closeness test $\Delta^{N(\varphi)}$ given by*

$$\frac{N(\varphi(\overline{X}_1, \overline{Y}))}{N(\varphi(\overline{X}_1\overline{Y}))} \overset{1}{\underset{2}{\gtrless}} \frac{N(\varphi(\overline{X}_2, \overline{Y}))}{N(\varphi(\overline{X}_2\overline{Y}))}$$

has error probability $P_{\mathrm{e}}(\Gamma^{\Delta^{N(\varphi)}}, P_1, P_2) \le \sqrt{\delta} \cdot e^{7n^{2/3}}$.

Proof Sketch. When $\overline{X}_1, \overline{X}_2, \overline{Y} \sim P_1^n \times P_2^n \times P_1^n$, since (P_1, P_2) are (n, δ)-different, by Theorem 23,

$$\Pr\Big(\frac{N(\varphi(\overline{X}_1, \overline{Y}))}{N(\varphi(\overline{X}_1\overline{Y}))} \le \sqrt{\delta} \cdot \sqrt[4]{\pi n} e^{\frac{1}{12n}}\Big) \le \sqrt{\delta} \cdot e^{6n^{2/3}} \sqrt{\pi n} e^{\frac{1}{6n}}$$

and

$$\Pr\Big(\frac{N(\varphi(\overline{X}_2, \overline{Y}))}{N(\varphi(\overline{X}_2\overline{Y}))} > \sqrt{\delta} \cdot \sqrt[4]{\pi n} e^{\frac{1}{12n}}\Big) \le \sqrt{\delta} \cdot e^{6n^{2/3}} \sqrt{\pi n} e^{\frac{1}{6n}}.$$

Hence, by union bound,

$$\begin{aligned} P_{1,2,1}\big(\Gamma^{\Delta^{N(\varphi)}}(\overline{X}_1, \overline{X}_2, \overline{Y}) = 2\big) &= \Pr\Big(\frac{N(\varphi(\overline{X}_1, \overline{Y}))}{N(\varphi(\overline{X}_1\overline{Y}))} \le \frac{N(\varphi(\overline{X}_2, \overline{Y}))}{N(\varphi(\overline{X}_2\overline{Y}))}\Big) \\ &\le 2\sqrt{\delta} \cdot e^{6n^{2/3}} \sqrt{\pi n} e^{\frac{1}{6n}}. \end{aligned}$$

Similar is the case when $\overline{Y} \sim P_2^n$ and the result follows. □

It is also worth noting that most of the known classifiers are essentially based on *relative* tests for closeness between $(\overline{X}_1, \overline{Y})$ and $(\overline{X}_2, \overline{Y})$. In the next section we consider a direct approach to classification.

4.2 Classifiers based on direct GLRT on profiles

Since classification is a composite hypothesis testing problem between the classes $\mathcal{P}_1$ and $\mathcal{P}_2$ defined earlier, it is natural to consider GLRTs based on these classes. We thus have the following GLRTs and their error guarantees.

Lemma 34. *For all distribution pairs P_1, P_2 that are (n,δ)-classifiable, the classifier $\Gamma^{\hat{P}(\varphi)}$ given by*

$$\hat{P}_{1,2,1}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y})) \overset{1}{\underset{2}{\gtrless}} \hat{P}_{1,2,2}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))$$

has error probability $P_{\mathrm{e}}(\Gamma^{\hat{P}(\varphi)}, P_1, P_2) \le \delta \cdot e^{8n^{3/4}}$. Here $\hat{P}_{1,2,1}$ corresponds to the maximum likelihood over $\mathcal{P}_1$ and $\hat{P}_{1,2,2}$ over $\mathcal{P}_2$.

Proof Sketch. Let $\mathcal{P}_1' \overset{\text{def}}{=} \{P_{3,4,3}(\Phi^{n,n,n}) : P_3, P_4$ are (n,δ)-classifiable$\}$ and similarly $\mathcal{P}_2' \overset{\text{def}}{=} \{P_{3,4,4}(\Phi^{n,n,n}) : P_3, P_4$ are (n,δ)-classifiable$\}$ be the (n,δ)-classifiable subsets of $\mathcal{P}_1$ and $\mathcal{P}_2$. Also let $\hat{P}'_{1,2,1}(\overline{\overline{\varphi}})$ and $\hat{P}'_{1,2,2}(\overline{\overline{\varphi}})$ denote the maximum likelihood of $\overline{\overline{\varphi}}$ under $\mathcal{P}_1'$ and $\mathcal{P}_2'$ respectively. Let $\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}) = \overline{\overline{\varphi}}$. If $\overline{Y} \sim P_1^n$, by using P1 of Lemma 19, and the fact that P_1, P_2 is (n,δ)-classifiable, $\hat{P}'_{1,2,1}(\overline{\overline{\varphi}}) > \hat{P}_{1,2,2}(\overline{\overline{\varphi}})$ with error probability at most $\delta \cdot |\Phi^{n,n,n}|$. Since $\hat{P}_{1,2,2}(\overline{\overline{\varphi}}) \ge \hat{P}'_{1,2,1}(\overline{\overline{\varphi}})$, it follows that $P_{1,2,1}\big(\Gamma^{\Delta^{N(\varphi)}}(\overline{X}_1, \overline{X}_2, \overline{Y}) = 2\big) \le \delta \cdot |\Phi^{n,n,n}|$. Similar is the case when $\overline{Y} \sim P_2^n$. Thus, the result follows using the bound on $|\Phi^{n,n,n}|$ from Lemma 2. □

Similar to the problem of closeness testing, by replacing the maximum likelihoods of the profiles in the GLRT with their approximations based on their inverse pattern counts, we obtain the following classifier and along with its error guarantee.

Theorem 35. *For all distribution pairs P_1, P_2 that are (n,δ)-classifiable, the classifier $\Gamma^{N(\varphi)}$ given by*

$$N(\varphi(\overline{X}_1, \overline{X}_2\overline{Y})) \underset{2}{\overset{1}{\gtrless}} N(\varphi(\overline{X}_1\overline{Y}, \overline{X}_2))$$

has error probability $P_{\mathrm{e}}(\Gamma^{N(\varphi)}, P_1, P_2) \le \sqrt{\delta} \cdot e^{18n^{3/4}} + e^{-n^{3/4}}$.

Proof Sketch. Along the lines of Observation 21 and Lemma 22, it can be shown that

$$\hat{P}_{1,2,1}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y})) \ \le\ \frac{N(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))}{N(\varphi(\overline{X}_1\overline{Y}, \overline{X}_2))} \ \le\ \frac{\hat{P}_{1,2,1}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))}{\hat{P}_{1,2,3}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))} \cdot \sqrt{\pi n} e^{\frac{1}{6n}}$$

and

$$\hat{P}_{1,2,2}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y})) \ \le\ \frac{N(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))}{N(\varphi(\overline{X}_1, \overline{X}_2\overline{Y}))} \ \le\ \frac{\hat{P}_{1,2,2}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))}{\hat{P}_{1,2,3}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))} \cdot \sqrt{\pi n} e^{\frac{1}{6n}}.$$

Hence, we have

$$\begin{aligned} &\frac{\hat{P}_{1,2,3}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))}{\sqrt{\pi n} e^{\frac{1}{6n}}} \frac{\hat{P}_{1,2,1}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))}{\hat{P}_{1,2,2}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))} \\ &\qquad \le\ \frac{N(\varphi(\overline{X}_1, \overline{X}_2\overline{Y}))}{N(\varphi(\overline{X}_1\overline{Y}, \overline{X}_2))} \ \le \\ &\qquad \frac{\hat{P}_{1,2,1}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))}{\hat{P}_{1,2,2}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))} \frac{\sqrt{\pi n} e^{\frac{1}{6n}}}{\hat{P}_{1,2,3}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y}))}. \end{aligned}$$

Let $\overline{\overline{\varphi}} = \varphi(\overline{X}_1, \overline{X}_2, \overline{Y})$. When $\overline{Y} \sim P_1^n$,

$$\begin{aligned} P_{1,2,1}\big(\hat{P}_{1,2,3}(\overline{\overline{\varphi}})) \le e^{-9n^{3/4}}\big) &\le P_{1,2,1}\big(P_{1,2,1}(\varphi(\overline{X}_1, \overline{X}_2, \overline{Y})\big) \\ &\le |\Phi^{n,n,n}| \cdot e^{-9n^{3/4}}\big) \\ &\le e^{-n^{3/4}}. \end{aligned}$$

Thus, with high probability $\ge 1 - e^{-n^{3/4}}$, $\sqrt{\pi n} e^{\frac{1}{6n}} / \hat{P}_{1,2,3}(\overline{\overline{\varphi}}) \le e^{10n^{3/4}}$. When this bound holds, using P6 of Lemma 19, (and considering $\mathcal{P}_1'$ and $\mathcal{P}_2'$ similar to Theorem 34), it follows that the error probability is $\sqrt{\delta} \cdot e^{18n^{2/3}}$. Using union bound on the error probabilities, and performing a similar analysis for $\overline{Y} \sim P_2^n$, the result follows. □

Variants of a different but simple classifier $\Gamma^{\mathrm{Q}(\Psi)}$ obtained as a straightforward application of the pattern probability estimators in [49] were considered in [58, 4]. It is motivated by the fact that if a classifier has access only to $\mathcal{M}(P_1)$ and $\mathcal{M}(P_2)$ (but not $\mathcal{M}(P_1, P_2)$), the test obtained by assuming a uniform prior over all possible mappings of $\mathcal{M}(P_1)$ and $\mathcal{M}(P_2)$ to the underlying alphabet $\mathcal{A}$ is given by

$$\frac{1}{(k-m_1)^{\underline{\Delta m_1}}}\frac{P_1(\Psi(\overline{X}_1\overline{Y}))}{P_1(\Psi(\overline{X}_1))} \underset{2}{\overset{1}{\gtrless}} \frac{1}{(k-m_2)^{\underline{\Delta m_2}}}\frac{P_2(\Psi(\overline{X}_2\overline{Y}))}{P_2(\Psi(\overline{X}_2))},$$

where $m_1 = m(\overline{X}_1)$, $\Delta m_1 = m(\overline{X}_1\overline{Y}) - m(\overline{X}_1)$, and m_2 and Δm_2 are defined similarly. This is followed by replacing the pattern probabilities by their maximum likelihood counterparts, and thereafter by their approximations using pattern counts of their profiles. In the next section, experimental results for text categorization are shown for a smoothed version of this estimator, which is based on the computationally efficient sequential pattern probability estimator in [49]. It is referred to as the single pattern (SP) classifier in the experiments. Experimental results are also shown for a similarly smoothed version of $\Gamma^{N(\varphi)}$, and is referred to as the joint pattern (JP) classifier. The classifier $\Gamma^{\Delta^{N(\varphi)}}$ is similar to $\Gamma^{\mathrm{Q}(\Psi)}$ and is not considered in the experiments.

4.3 Text classification experiments

We show experimental results for text classification to demonstrate the performance of pattern based classifiers. In this application, one is given a data set consisting of documents, for example, electronic messages from newsgroups, along with their pre-assigned labels, for example, their topic, and the task is to label new documents.

One of the techniques that works reasonably well in practice is *Naive Bayes* [39], which assumes a *Bag of Words* model, *i.e.,* the words in each document are generated *i.i.d.* according to the distribution of the class to which it belongs. Naive Bayes classifiers are LRT's that use one of the several well known probability estimators, for example, Laplace or Good-Turing estimators, to estimate the underlying distributions of the classes from the training documents. Our experiments show that pattern based classifiers, which are essentially Naive Bayes classifiers that use pattern probability estimators, can perform as good as the state-of-the-art techniques like Support Vector Machine (SVM).

We use the `rainbow` toolkit [43] for classification, with additional support for pattern based classifiers and optimal classifiers that use actual distributions for synthetic data sets. We compare between Laplace (`lap`), SVM with linear kernel (`svm`), and the classifiers based on single patterns (`sp`) and joint patterns (`jp`) described at the end of Section 4.2. We note that the SVM with linear kernel with word counts as features is equivalent to the classifier analyzed in [35] that looks at the L_2 distance between the empirical distributions of the training and test sequences.

4.3.1 Synthetic data sets

These experiments are intended to demonstrate that pattern based classifiers work well when the data sets indeed confirm to the Bag of words model. The data sets, which try to resemble actual data sets, were generated as follows. For simplicity we consider data sets with 2 classes. The distributions corresponding to both classes have the same monotone distribution, *i.e.,* probability multiset, which is a Zipf distribution [76]:

$$p_i = \frac{c}{(i_0 + i)^e},$$

for $i = 1, 2, \ldots, k$. The exponent e is usually 1 or close to 1. The normalizing constant is c and the initial offset is i_0. In the experimental results shown in Table 4.1, $k = 30,000$, $i_0 = 500$ and exponent e takes values 1 or 0.8. The actual distributions P_1 and P_2 is obtained by permuting the monotone distribution, and ensuring that the two distributions are not too different so that they are non-trivial to classify, similar to real data sets. This is achieved by permuting the probabilities randomly such that the final index of each probability is within a range that is grows with the original index. Specifically, if a word has rank i_1, *i.e.,* probability p_{i_1} in P_1, then its rank in P_2 is $i_2 = \texttt{rank}(i_1 + R \cdot i_1^x)$ (and probability in P_2 is p_{i_2}) where R is a uniform random number in the range $[0, 1]$ and the power x, typically between 1 and 2, determines how far the indices can get permuted from their original values. Thus, smaller probabilities at the tail of the distribution are permuted within a farther range. There are 1000 documents per class and 75 words per document and the documents are split 50-50 into training and test. Thus, the length of training sequences is $1000 \times 75 \times 0.5 = 37500$. And there are 500 test sequences of length 75 from each class. The error probability is the average over these 1000 test sequences. We also average over multiple random splits and multiple random index permutations. It can be seen from the results shown in Table 4.1, that pattern based classifiers, particularly joint patterns, perform favorably.

Table 4.1: Accuracy of different classifiers on synthetic datasets.

e	x	lap	svm	sp	jp
1.0	1.0	86.6	88.0	86.0	87.2
1.0	1.3	93.5	90.7	94.0	94.1
1.0	1.6	88.8	82.5	89.1	90.0
0.8	1.2	85.8	86.8	88.7	87.2
0.8	1.5	90.7	89.9	92.5	92.6
0.8	1.8	93.8	92.1	94.9	94.9

4.3.2 Real world data sets

These experiments demonstrate the favorable performance of pattern based classifiers on some of the well known actual data sets. The collection *Newsgroups*, *i.e.,* 20ng, is a list of 1000 articles collected from 20 newsgroups. It contains several closely related subgroups, for example, comp.*, sci.* and talk.*. The *Reuters 21758* data sets, *i.e.,* r52 and a subset r8, have 52 and 8 classes respectively and the number of documents per class vary sharply between few thousands to just one or two. The *CADE* dataset, *i.e.,* cade, is a collection of Portuguese web documents consisting of 12 classes. It is a fairly large and uneven data set with documents per class ranging between few hundreds to few thousands and is generally difficult to classify. The data set *World Wide Knowledge Base*, *i.e.,* webkb, is a small data set of 4 classes of variable number of documents per class. These data sets, along with their training-test split can be obtained from [12]. The results are shown in Table 4.2. While results in general are in favor of SVM, they also show the favorable performance of pattern based classifiers.

Table 4.2: Accuracy of different classifiers for real data sets.

Data set	Classification method			
	lap	svm	sp	jp
webkb	83.30	87.94	83.37	83.41
20ng	80.76	80.80	82.68	83.31
r52	80.53	92.04	85.59	89.18
cade	53.10	52.09	57.01	55.96

Acknowledgement

Chapter 4 is adapted from Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, Narayana Prasad Santhanam, "Classification using pattern maximum likelihood", *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 2010. The dissertation author was the primary researcher and author of this paper.

Chapter 5

Distribution Multiset Estimation

In this chapter, we consider the problem of explicitly estimating the multiset of probability values of a discrete distribution given *i.i.d.* samples from it. Specifically, let $\mathcal{A} \stackrel{\text{def}}{=} \{a_1, \ldots, a_k\}$ denote the alphabet as earlier and $P = (P(a_1), \ldots, P(a_k))$ be a probability distribution on $\mathcal{A}$. We recall from Section 2.3 that the probability multiset of P is the collection of probability values in P and denoted by $\mathcal{M}(P) \stackrel{\text{def}}{=} (p_1, \ldots, p_k) \stackrel{\text{def}}{=} \{P(a_1), \ldots, P(a_k)\}$ where $p_1 \geq p_2 \geq \cdots \geq p_k$. Given a length-$n$ *i.i.d.* sequence $\overline{X} \sim P^n$ we want to estimate the probability multiset $\mathcal{M}(P)$. An estimator $Q \stackrel{\text{def}}{=} Q_{\overline{X}} = (q_1, \ldots, q_k)$ outputs a probability multiset corresponding to each input sequence $\overline{X} \in \mathcal{A}^n$. The estimator need not provide estimates $Q(a)$ for the probabilities of specific symbols $a \in \mathcal{A}$. We want to obtain Q such that for some suitable distance metric $D(\cdot, \cdot)$ defined on distribution multisets and some $\epsilon > 0$, $D(P, Q) \leq \epsilon$ with high probability $1 - o_n(1)$.

One such distance that is commonly considered is the (sorted) L_1 distance $L_1(P, Q) = \sum_{i=1}^{k} |p_i - q_i|$ where $p_1 \geq p_2 \geq \cdots p_k$ and $q_1 \geq q_2 \geq \cdots q_k$. We use $|P - Q|_1$ to specifically denote the L_1 distance on probability multisets instead of $|P - Q|$, which is used in earlier chapters with the definition $\sum_i |P(a_i) - Q(a_i)|$. Note that among all permutations $\sigma : [k] \to [k]$, $|P - Q|_1 = \min_{\sigma \in S_k} |P - \sigma(Q)| = \min_\sigma \sum_i |P(a_i) - Q(a_{\sigma(i)})|$. Another distance that has been used recently in [68] is the relative earthmover distance $R(P, Q)$. In can be described as the minimum cost of moving the probability mass of P to make it equal to Q, where the per unit cost of moving mass from p_i to q_j is $|\log \frac{p_i}{q_j}|$. The following fact relating $R(P, Q)$

and $|P-Q|_1$ is useful.

Fact 36. *For all P and Q, $\frac{1}{2}|P-Q|_1 \le R(P,Q)$.* □

These distances can be motivated by their relevance to property estimation and testing, *e.g.,* as shown in [70, 68], due to a (ϵ,δ)-continuity relationship between various properties π and these distances $D(\cdot,\cdot)$ of the form given by "$D(P,Q) \le \epsilon$ implies $|\pi(P)-\pi(Q)| \le \delta$". For example, for the case of entropy, $|H(P)-H(Q)| \le |P-Q|_1 \log_2(\frac{k}{|P-Q|_1})$ and $|H(P)-H(Q)| \le R(P,Q)$. Thus, obtaining an estimate Q such that $D(P,Q) \le \epsilon$ automatically implies $\phi = \pi(Q)$ satisfies $|\phi - \pi(P)| \le \delta$. (We assumed here that the properties are real. As such we can consider the multiset itself as the property and thus consider continuity between distances. For example, $R(P,Q) \le \epsilon$ implies $|P-Q|_1 \le \epsilon$.)

We further observe that if an estimator Q approximates P to within a small distance with error probability $< \frac{1}{2}$, then it can be improved to any $\delta > 0$ using sequences of length $n = \mathcal{O}(\log(\frac{1}{\delta}))$. This is shown in the Observation below.

Observation 37. *Let $D(\cdot,\cdot)$ be a distance metric on distributions, and let Q be an estimator for a collection of distributions $\mathcal{P}$ such that for all $P \in \mathcal{P}$, $\Pr(D(P,Q_{\overline{X}}) \ge \epsilon) \le \delta \le \frac{1}{4}$ when given a length-n sequence $\overline{X} \sim P^n$. Then, for all positive integers r, there is an estimator Q' that for all $P \in \mathcal{P}$, $\Pr(D(P,Q'_{\overline{X}'}) \ge 3\epsilon) \le (4\delta)^r$ when given a length-n' sequence $\overline{X}' \sim P^{n'}$ where $n' = (2r+1)n$.*

Proof. For any $P \in \mathcal{P}$, given a sequence of $\overline{X}' \sim P^{n'}$ of length $n' = (2r+1)n$, divide it into $(2r+1)$ equal parts of length n. Let $\{Q_1, Q_2, \ldots, Q_{2r+1}\}$ be the output of Q on each of these $(2r+1)$ sequences of length n. (Note that all the $2r+1$ instances are independent.) Then, the probability that at least $r+1$ of these Q_i's are more than a distance of ϵ from Q is $\delta' = \sum_{j=2r+1}^{n} \binom{2r+1}{j} \delta^j (1-\delta)^{2r+1-j} \le r\binom{2r+1}{r}\delta^r \le (4\delta)^r$. Hence with high probability $1-(4\delta)^r$, there are at least $(2r+1)$ Q_i's within a distance of ϵ from P, and therefore at a distance of within 2ϵ from each other by triangle inequality. Therefore, with error probability $(4\delta)^r$, there exists at least one "clique" of more than $(2r+1)$ Q_i's that are within a distance of 2ϵ from each other. We find one such set of Q_i's and output any Q_i in that group as the estimate $Q'_{\overline{X}'}$. Notice that there may be several such sets of which at least one set has all

Q_i's that are within 2ϵ from P. Furthermore, any two sets of size $r+1$ have one Q_i in common, and thus by triangle inequality, Q' is within $2\epsilon + \epsilon = 3\epsilon$ away from P. (The distance between Q' and a common "correct" Q_i is 2ϵ, and this correct Q_i is within ϵ from P.) □

Similar to the case of closeness testing, it is worthwhile to consider the empirical distribution estimator, $Q^{\text{emp}}_{\overline{X}} \overset{\text{def}}{=} \tau(\overline{X})$, *i.e.*, $\{\frac{\mu_i}{n} : i = 1, 2, \ldots, m\}$, following the notation in Section 2.2. We consider its estimation properties in terms of L_1 distance. While this result can also be shown using simple applications of Chernoff bounds, we take a different approach with some more insights.

Lemma 38. *For all sufficiently large n and $\epsilon > 0$, for all distributions P whose support size is $k = \mathcal{O}(\epsilon^{2.1}n)$, given $\overline{X} \sim P^n$,* $\Pr(|P - Q^{\text{emp}}_{\overline{X}}|_1 \geq \epsilon) \leq e^{-n\epsilon^2/8}$. *Furthermore,* $\Pr(|P - Q^{\text{emp}}_{\overline{X}}| \geq \epsilon) \leq e^{-n\epsilon^2/8}$ *and* $\Pr(D(Q^{\text{emp}}_{\overline{X}}||P) \geq \epsilon) \leq e^{-n\epsilon}$, *where* $D(P||P') \overset{\text{def}}{=} \sum_{a\in\mathcal{A}} P(a)\log(\frac{P(a)}{P'(a)})$.

Proof. For all sequences $\overline{X} \in \mathcal{A}^n$, we have

$$\frac{P(\overline{X})}{Q^{\text{emp}}_{\overline{X}}(\overline{X})} = \frac{\prod_{a\in\mathcal{A}} P(a)^{\mu(a)}}{\prod_{a\in\mathcal{A}} (\frac{\mu(a)}{n})^{\mu(a)}} = e^{-nD(Q^{\text{emp}}_{\overline{X}}||P)}.$$

If $|P - Q|_1 \geq \epsilon$, then $D(Q||P) \geq \frac{1}{2}|P - Q|^2 \geq \frac{1}{2}|P - Q|_1^2 \geq \frac{1}{2}\epsilon^2$. Thus,

$$\begin{aligned}
\Pr(|P - Q^{\text{emp}}_{\overline{X}}| \geq \epsilon) &= \sum_{\overline{x}:|P-Q^{\text{emp}}_{\overline{x}}|\geq\epsilon} P(\overline{x}) \\
&= \sum_{\overline{x}:|P-Q^{\text{emp}}_{\overline{x}}|\geq\epsilon} Q^{\text{emp}}_{\overline{x}}(\overline{x}) \cdot e^{-nD(Q^{\text{emp}}_{\overline{x}}||P)} \\
&\leq \sum_{\overline{x}:|P-Q^{\text{emp}}_{\overline{x}}|\geq\epsilon} Q^{\text{emp}}_{\overline{x}}(\overline{x}) \cdot e^{-\frac{1}{2}n\epsilon^2} \\
&\leq e^{-\frac{1}{2}n\epsilon^2} \sum_{\overline{x}} Q^{\text{emp}}_{\overline{x}}(\overline{x}).
\end{aligned}$$

The quantity $A(n,k) \overset{\text{def}}{=} \sum_{\overline{x}} Q^{\text{emp}}_{\overline{x}}(\overline{x}) = \sum_{\mu_1+\cdots+\mu_k=n} \binom{n}{\mu_1,\ldots,\mu_k} \prod_{i=1}^k (\frac{\mu_i}{n})^{\mu_i}$ (where the summation is over ordered k-tuples $(\mu_1, \ldots, \mu_k)$), known as *Shtarkov sum* for *i.i.d.* sequences of length-n and alphabet size k has been analyzed extensively in the context of universal compression [20, 59, 62, 73]. It has been shown in [47] that when $k = o(n)$, then $\log(A(n,k)) = \frac{k-1}{2}\log\frac{n}{k}(1+o(1)) = o(n)$, and when $k = \Theta(n)$,

$\log(A(n,k)) = \Theta(n)$. The precise asymptotics have been considered recently in [64, 65] and it has been shown that for $\alpha \to 0$, and $k = \alpha n$, $\log(A(n,k)) \approx (\frac{\alpha}{2}\log\frac{1}{\alpha})n$. Hence, for small ϵ, setting $\alpha = \frac{\epsilon^{2.1}}{2}$, we get the desired result. The result in terms of the other two distances follow similarly. □

Since Q^{emp} guarantees that $|Q^{\text{emp}}_{\overline{X}} - P|$ is small and not just $|Q^{\text{emp}}_{\overline{X}} - P|_1$, this can be used for the closeness testing problem considered in Chapter 3.

Observation 39. *For all sufficiently large n and any $\epsilon > 0$, given $(\overline{X}_1, \overline{X}_2) \sim P_1^n \times P_2^n$, the closeness test $|\tau(\overline{X}_1) - \tau(\overline{X}_2)| \underset{diff}{\overset{same}{\lessgtr}} \frac{\epsilon}{2}$ has error probability at most $2e^{-n\epsilon^2/8}$ whenever $P_1 = P_2$ or $|P_1 - P_2| > \epsilon$.* □

For examples where $k = \Theta(n)$ and this estimator does not perform well, *i.e.,* $|P - Q^{\text{emp}}_{\overline{X}}|$ is large with high probability and cannot be used for closeness testing or classification, see [35]. It is evident from the above lemma and observation that Q^{emp} not only estimates $\mathcal{M}(P)$ but also the probabilities of specific symbols. Indeed it is well known that estimators Q for distributions P (not just the multiset) such that $|P - Q| = o_n(1)$ are possible only when $k = o(n)$. The problem of estimating $\mathcal{M}(P)$ has a much weaker requirement (say in terms of sorted vs. unsorted L_1 distance, because $|P - Q|_1 \le |P - Q|$), so intuitively, we should be able to estimate well for even larger k, although the exact limits are not clear. It has however been shown in [70] that estimating $\mathcal{M}(P)$ in L_1 distance still requires $n \ge k/2^{\Theta(\sqrt{\log(k)})}$ samples, or equivalently distributions of support size $k \le n \cdot 2^{\Theta(\sqrt{\log(n)})}$ (which is $o(n^{1+\epsilon})$ for any $\epsilon > 0$). Valiant *et al.* in [68] strengthen these bounds. They show a computationally efficient estimator that approximates the distributions to within an earthmover distance of ϵ using n samples, whenever the alphabet size is $k = \mathcal{O}(\epsilon^{2.1} n \log(n))$, and show matching upper bounds (*i.e.,* lower bounds in terms of sample complexity).

In the next section, we analyze the estimation properties of the profile maximum likelihood estimator $Q^{\text{PML}}_{\overline{X}} \overset{\text{def}}{=} \hat{P}_{\varphi(\overline{X})} = \arg\max_P P(\varphi(\overline{X}))$. We show that they are competitive with respect to any estimator in any distance metric. Furthermore, in its most general form, it does not assume any bounds on the alphabet size. We again do this via a general competitive property of ML distribution estimator which is similar in flavor to the result we showed for composite hypothesis testing in Section 3.3.

5.1 Competitivity of the PML estimator

5.1.1 Competitivity of ML for distribution estimation

We show a general fact about the competitive optimality of maximum likelihood estimators. Let $\mathcal{Z}$ be a discrete alphabet of size $|\mathcal{Z}|$ and $\mathcal{P}$ be a collection of probability distributions on $\mathcal{A}$. Given a sample Z generated according to an unknown distribution $P \in \mathcal{P}$, we want to estimate P. An estimator $Q : \mathcal{Z} \to \mathcal{P}$, outputs a distribution $Q_z \in \mathcal{P}$ corresponding to a given sample $z \in \mathcal{Z}$. The maximum likelihood (ML) estimator outputs a distribution $\hat{P}_z \in \mathcal{P}$ that maximizes the likelihood of observing z, *i.e.,*

$$\hat{P}_z \stackrel{\text{def}}{=} \hat{P}_{\mathcal{P},z} \stackrel{\text{def}}{=} \arg\max_{P\in\mathcal{P}} P(z).$$

We also use

$$\hat{P}(z) \stackrel{\text{def}}{=} \hat{P}_{\mathcal{P}}(z) \stackrel{\text{def}}{=} \max_{P\in\mathcal{P}} P(z) \tag{5.1}$$

to denote the maximum likelihood of z under any distribution in $\mathcal{P}$.

To measure how good an estimate is, let $D : \mathcal{P} \times \mathcal{P} \to \mathbb{R}^{\geq 0}$ be a distance defined on distributions in $\mathcal{P}$ that is a metric, in particular, nonnegative, symmetric and satisfies the triangle inequality, *i.e.,* $D(P, P') = D(P', P)$ and $D(P, P') \leq D(P, P'') + D(P'', P')$ for all distributions $P, P', P'' \in \mathcal{P}$. We say that an estimator Q is a (ϵ, δ)-good estimator of a distribution P with respect to distance $D(\cdot, \cdot)$ for some $\delta \in [0, 1]$ and $\epsilon \geq 0$, if given $Z \sim P$,

$$\Pr(D(P, Q_Z) \geq \epsilon) = \sum_{z\in\mathcal{Z}:D(P,Q_z)\geq\epsilon} P(z) \leq \delta.$$

The next lemma shows that $\hat{P}_Z$ is as good as any other estimator.

Lemma 40. *Let Q be an estimator such that for all $P \in \mathcal{P}$, given $Z \sim P$,*

$$\Pr\left(D(P, Q_X) \geq \epsilon\right) \leq \delta \tag{5.2}$$

for some fixed $\epsilon \geq 0$ and $\delta \in [0,1]$. Then,

$$\Pr\left(D(P, \hat{P}_X) \geq 2\epsilon\right) \leq \delta \cdot |\mathcal{Z}|.$$

In other words, if there exists an (ϵ, δ)-good estimator Q for all distributions $P \in \mathcal{P}$, then the ML estimator $\hat{P}$ is also a $(2\epsilon, \delta|\mathcal{Z}|)$-good estimator.

Proof. We consider separately the cases when the generated symbol $Z = z$ is such that $P(z) > \delta$ and when $P(z) \leq \delta$. When $P(z) > \delta$, we make an easy claim that $D(P, \hat{P}_z) \leq 2\epsilon$. To see this, we observe that clearly $D(P, Q_z) \leq \epsilon$, otherwise, if $D(P, Q_z) \geq \epsilon$, then

$$\Pr\left(D(P, Q_Z) \geq \epsilon\right) \; = \sum_{z \in \mathcal{Z}: D(p, q_z) \geq 2\epsilon} P(z) \; \geq \; P(z) \; > \; \delta,$$

contradicting that Q is a (ϵ, δ)-good estimator of P. By a similar reasoning, $D(\hat{P}_z, Q_z) \leq \epsilon$, since Q is a (ϵ, δ)-good estimator of $P' = \hat{P}_z \in \mathcal{P}$ and $P'(z) = \hat{P}(z) \geq P(z) > \delta$ as well. Hence, $D(P, \hat{P}_z) \leq D(P, Q_z) + D(Q_z, \hat{P}_z) \leq 2\epsilon$, proving the claim.

For the case when $P(z) \leq \delta$, we note that

$$\Pr\left(P(Z) \leq \delta\right) \; = \sum_{z \in \mathcal{Z}: P(z) \leq \delta} P(z) \; \leq \; \delta \cdot |\mathcal{Z}|.$$

Combining the two cases,

$$\begin{aligned} \Pr\left(D(P, \hat{P}_Z) \geq 2\epsilon\right) \; &= \; \Pr\left((D(P, \hat{P}_Z) \geq 2\epsilon) \wedge (P(Z) > \delta)\right) \\ &\qquad + \Pr\left((D(P, \hat{P}_Z) \geq 2\epsilon) \wedge (P(Z) \leq \delta)\right) \\ &\leq \; 0 + \Pr(P(Z) \leq \delta) \\ &\leq \; \delta|\mathcal{Z}|. \end{aligned}$$

□

5.1.2 PML for distribution multiset estimation

Clearly, the probability multiset estimation problem is a special case of the above general distribution estimation. Since we want to estimate $\mathcal{M}(P)$ using $\overline{X} \sim P^n$, without loss of generality, we restrict ourselves to profile-based estimators that depend on $\overline{X}$ through its $\varphi(\overline{X})$, *e.g.*, see [9, Section 3.1.3] for a simple argument or by using a similar argument provided for closeness testing in Subsection 3.2.1. Then, we have the correspondence with the general distribution estimation problem that $\mathcal{Z} \leftrightarrow \Phi^n$, $\mathcal{P}$ consists of all *i.i.d.* profile distributions $P(\Phi^n)$, and $D(\cdot, \cdot)$ is any distance defined on probability multisets. For distribution estimation, we say that an estimator Q is an (n, ϵ, δ)-good estimator for a class of distributions $\mathcal{P}$ if for all $P \in \mathcal{P}$, $\Pr(|P - Q_{\overline{X}}| \geq \epsilon) \leq \delta$ given $\overline{X} \sim P^n$. For a class of distributions $\mathcal{P}$ and for all profiles $\overline{\varphi} \in \Phi^n$, we denote the class restricted PML distribution and the corresponding likelihood as

$$\hat{P}_{\mathcal{P},\overline{\varphi}} \stackrel{\text{def}}{=} \arg\max_{P \in \mathcal{P}} P(\overline{\varphi}) \quad \text{and} \quad \hat{P}_{\mathcal{P}}(\overline{\varphi}) \stackrel{\text{def}}{=} \max_{P \in \mathcal{P}} P(\overline{\varphi}) = \hat{P}_{\mathcal{P},\overline{\varphi}}(\overline{\varphi}).$$

In particular, we use the notation $\mathcal{P}_k$ to denote the class of all distributions of support size k. We use $\hat{P}_k(\overline{\varphi})$ and $\hat{P}_{k,\overline{\varphi}}$ to denote the maximum likelihood of $\overline{\varphi}$ and maximizing distribution under $\mathcal{P}_k$.

Lemma 41. *Let $\mathcal{P}$ be a class of distributions for which there exists a profile-based probability multiset estimator $Q_{\varphi(\overline{X})}$ such that for some distance $D(\cdot, \cdot)$ defined on distribution multisets, and some ϵ, δ and for all $P \in \mathcal{P}$, when $\overline{X} \sim P^n$,*

$$\Pr\left(D(P, Q_{\varphi(\overline{X})}) \geq \epsilon\right) \leq \delta.$$

Then, the PML distribution multiset estimator has error

$$\Pr\left(D(P, \hat{P}_{\mathcal{P},\varphi(vecX)}) \geq 2\epsilon\right) \leq \delta \cdot |\Phi^n| \leq \delta \cdot e^{3n^{1/2}}.$$

In other words, if there is a (n, ϵ, δ)-good estimator $Q_{\varphi(\overline{X})}$ for distributions in $\mathcal{P}$, then $\hat{P}_{\mathcal{P},\overline{\varphi}}$ is a $(n, 2\epsilon, \delta|\Phi^n|)$-good and hence $(n, 2\epsilon, \delta \cdot e^{3\sqrt{n}})$-good estimator for $\mathcal{P}$.

Proof. Using the correspondence with the general distribution estimation problem, the result follows from Lemma 40 along with Lemma 1, which states that $|\Phi^n| \leq e^{\pi\sqrt{\frac{2}{3}}\sqrt{n}} < e^{3\sqrt{n}}$. □

This along with Observation 37 implies the following estimation guarantee for PML estimator in terms of sample complexity.

Corollary 42. *Following the setup in Lemma 41,* i.e., *if there is an estimator $Q_{\varphi(\overline{X})}$ for $\mathcal{P}$ such that for all $P \in \mathcal{P}$, when $\overline{X} \sim P^n$,*

$$\Pr\left(D(P, Q_{\varphi(\overline{X})}) \geq \epsilon\right) \leq \delta < \frac{1}{4},$$

then there is an estimator $Q'_{\varphi(\overline{X}')}$ such that when given a sequence $\overline{X} \sim P^{n'}$ of length n',

$$\Pr\left(D(P, Q'_{\varphi(\overline{X}')}) \geq 3\epsilon\right) \leq \delta^2 \cdot e^{-6n'^{1/2}},$$

where $n' = \mathcal{O}\big(\max\{n, \frac{n^2}{\log^2(\frac{1}{4\delta})}\}\big)$. Thus, when given sequences $\overline{X}'$ of length n', the error probability of the PML estimator is

$$\Pr\left(D(P, \hat{P}_{\mathcal{P},\varphi(\overline{X}')}) \geq 6\epsilon\right) \leq \delta.$$

Proof Sketch. Using Observation 37, $r = \mathcal{O}\big(\max\{1, \frac{n}{\log^2(\frac{1}{4\delta})}\}\big)$ suffices to guarantee the existence of an estimator Q' whose error probability is $\delta' = (4\delta)^r \leq \delta^2 e^{-6\sqrt{n'}}$ using $\overline{X}' \sim P^{n'}$ where $n' = (2r+1)n$. Hence, by Lemma 41, the error probability of PML estimator is at most δ when given $\overline{X}' \sim P^{n'}$. □

Although the above sample complexity observation holds even when $\delta > e^{-3\sqrt{n}}$, Lemma 41 is largely useful when $\delta < e^{-3\sqrt{n}}$. Intuitively for most distances D, if $D(P,Q) = \epsilon > 0$, the affinity or overlap of their length-n profiles is $|P(\Phi^n) \wedge Q(\Phi^n)| < e^{-n \cdot f(\epsilon)}$ which essentially allows for existence of estimators with $\delta < e^{-n \cdot f(\epsilon)}$. This is indeed the case we see for sequence maximum likelihood in Lemma 38, and thus PML estimator $\hat{P}_{k,\varphi(\overline{X})}$ is also within L_1 distance of ϵ with high probability when $k = \mathcal{O}(\epsilon^{2.1} n)$. For stronger estimation guarantees, we consider the estimator by Valiant *et al.* in [68] which approximates distributions to within a relative earthmover distance of ϵ whenever $k = \mathcal{O}(\epsilon^{2.1} n \log(n))$ with low error probability $e^{-n^{0.03}}$. We show that the error probability can be improved to arbitrarily close to exponential, say $e^{-n^{0.9}}$, by minor modifications to the various constant parameters of their estimator (at the cost of much smaller constants in $k = \mathcal{O}(\epsilon^{2.1} n \log(n))$, thus again implying similar error guarantees for the PML (when restricted to $\mathcal{P}_k$). We briefly describe the modified estimator, which we henceforth refer to as Q^{VV}, and provide its error guarantee in the next subsection.

5.1.3 Valiants' estimator for superlinear alphabets

The following estimator, which we refer to as Q^{VV}, is considered in [68]. The version provided here only differs in the various constant parameters to improve upon the error probability.

The main ideas behind the estimator Q^{VV} are as follows. It is easier to motivate the estimator in the Poisson model where we are given sequences of length $n' \sim \text{poi}(n)$ instead of length n. See Section 2.6 for some of the preliminaries about the Poissonization technique. As noted in [68, 69] and in the following observation, both models are almost equivalent.

Observation 43. *For sufficiently large n and any $\epsilon > 0$ and any distance measure $D(\cdot,\cdot)$ on distributions, let Q be a distribution estimator for a class of distributions $\mathcal{P}$ such that for all $P \in \mathcal{P}$, when given input $\overline{X} \sim P^n$, $\Pr\left(D(P, Q_{\overline{X}}) \geq \epsilon\right) \leq \delta$. Then there is an estimator Q' that takes as input $\overline{X}' \sim P^{\text{poi}(n+n^{0.95})}$, and has error $\Pr\left(D(P, Q'_{\overline{X}'}) \geq \epsilon\right) \leq \delta + e^{-n^{0.91}}$.*

Similarly, if there is an estimator Q' that takes as input $\overline{X}' \sim P^{\text{poi}(n)}$ and has error $\Pr\left(D(P, Q'_{\overline{X}'}) \geq \epsilon\right) \leq \delta$, then there is an estimator Q' that takes as input $\overline{X} \sim P^{n+n^{0.95}}$ and has error $\Pr\left(D(P, Q_{\overline{X}}) \geq \epsilon\right) \leq \delta + e^{-n^{0.91}}$.

Proof. In the first case, consider a Q' outputs $Q_{\overline{X}''}$ where $\overline{X}''$ consists of first n samples of $\overline{X}'$ if length of $\overline{X}'$ is $n' \geq n$. By Poisson tail bounds of Observation 7, this happens with probability $\geq 1 - e^{-n^{0.91}}$. If $n' \leq n$, Q' outputs error, and thus has the stated error. A similar argument can be used for the second case. □

Suppose we are given a sequence $\overline{X}' \sim P^{\text{poi}(n)}$ where P is the unknown distribution we want to estimate. We immediately notice that probabilities in $\mathcal{M}(P)$ that are sufficiently high can be estimated accurately with their empirical frequencies. For any symbol a such that $P(a) \geq \frac{n^b}{n}$, where $b = 0.05$ is a small constant, since $\mu(a) \sim \text{poi}(nP(a))$, by Poisson tail bounds, $\mu(a)$ concentrates around its mean $nP(a)$ with high probability and thus $\frac{\mu(a)}{n}$ is a good estimate of $P(a)$. (This is similarly true when $\overline{X} \sim P^n$ by a simple application of Chernoff bounds.)

For estimating the low probabilities of $\mathcal{M}(P)$, *i.e.,* $\mathcal{M}_{\text{low}} = \{p : p \in \mathcal{M}(P), p \leq \frac{n^b}{n}\}$, we see that any such low probability symbol appears at most $2n^b$ times with high probability and thus contributes to the portion of the profile consisting of low multiplicity prevalences $(\varphi_1, \varphi_2, \ldots, \varphi_{2n^b})$. We thus need to estimate $\mathcal{M}_{\text{low}}$ only from these prevalences. To this end, we observe that prevalences closely concentrate around its mean as follows. Note that for any P,

$$\mathrm{E}_P[\varphi_\mu] \;=\; \mathrm{E}_P[\sum_{i=1}^{k} \mathbb{1}_{[\mu(a_i)=\mu]}] = \sum_{i=1}^{k} \mathrm{poi}(np_i, \mu).$$

Observation 44. *(Also [67, Corollary 22].) For all P, if* $\overline{X} \sim P^{\mathrm{poi}(n)}$ *and* $\varphi(\overline{X}) = (\varphi_1, \ldots, \varphi_n)$*, then for* $\mu = 1, 2, \ldots, n$ *and for all* $0.5 < \alpha < 1$*,*

$$\Pr\left(|\varphi_\mu - \mathrm{E}_P[\varphi_\mu]| \geq n^\alpha\right) \leq 2e^{-n^{2\alpha-1}/3}.$$

In particular, $\Pr\left(|\varphi_\mu - \mathrm{E}_P[\varphi_\mu]| \geq n^{0.99}\right) \leq e^{-n^{0.97}}$.

Proof Sketch. In the Poisson model, $\varphi_\mu = \sum_{a \in \mathcal{A}} \mathbb{1}_{[\mu(a)=\mu]}$ is a sum of independent 0-1 random variables, and thus the observation follows by Chernoff bounds. □

Hence, one would hope that if we find a Q such that $\mathrm{E}_Q[\varphi_\mu] \approx \varphi_\mu$, then $\mathrm{E}_Q[\varphi_\mu] \approx \varphi_\mu \approx \mathrm{E}_P[\varphi_\mu]$, and hence $Q(\Phi^n)$ and $P(\Phi^n)$ would be similar and that $\mathcal{M}(Q)$ and $\mathcal{M}(P)$ would also be similar. We would further hope that $\mathcal{M}_{\text{low}}$ can be approximated well by a Q whose probabilities take values among a fine enough grid $\{x_1, x_2, \ldots, x_\ell\}$, say $x_i = i/n^2$ for $i = 1, 2, \ldots, 2n^{1+b}$. Let $h_i \geq 0$ be the counts of x_i in a distribution Q. Then,

$$\mathrm{E}_Q[\varphi_\mu] = \sum_{i=1}^{\ell} h_i \cdot \mathrm{poi}(nx_i, \mu)$$

is linear in $(h_1, \ldots, h_\ell)$, for $\mu = 1, 2, \ldots, n$. Thus, such a Q that satisfies $\mathrm{E}_Q[\varphi_\mu] \approx \varphi_\mu$, say $|\mathrm{E}_Q[\varphi_\mu] - \varphi_\mu| \leq 2n^{0.99}$, for $\mu = 1, \ldots, 2n^b$, can be found by linear (integer) programming.

Combining the low probability estimates obtained from such a linear program along with the empirical estimates for high probabilities would result in an estimator for $\mathcal{M}(P)$. This is a basic reasoning for the estimator shown by

Valiant and Valiant in [67, 68]. Before proceeding to state the estimator, we define histograms h_P of distributions P, which are a convenient and equivalent way of representing $\mathcal{M}(P)$, just as profiles $\overline{\varphi}$ and multiplicity vectors $\overline{\mu}$ convey the same information.

Definition 45. (Also [67, Definition 4].) The histogram $h \overset{\text{def}}{=} h_P$ of a distribution P is a mapping $h : (0,1] \to \mathbb{R}$, where $h(x) = |\{i : p(i) = x, i \in \mathcal{A}\}|$. Generalized histograms are also allowed, that do not necessarily take integral values. □

The estimator Q^{VV} is defined as follows. Let $\overline{X}$ be the input sequence of length n and let $\overline{\varphi} = \varphi(\overline{X}) \in \Phi^n$. Let $a \overset{\text{def}}{=} 0.001$ be a small constant. One can always find $c \overset{\text{def}}{=} c_{\overline{\varphi}} \in [1,2]$ such that

$$\sum_{\mu=\lceil cn^a \rceil}^{\lceil cn^a + 4n^{0.6a} \rceil} \mu\varphi_\mu \ \le \ 4n^{1-0.4a}.$$

Clearly, such a c must exist, otherwise $\sum_{\mu=n^a}^{2n^a} \mu\varphi_\mu = \sum_{j=1}^{n^{0.4a}/4} \sum_{n^a+4(j-1)n^{0.6a}}^{n^a+4jn^{0.6a}} \mu\varphi_\mu > (n^{0.4a}/4) \cdot (4n^{1-0.4a}) = n$, which would contradict that $\sum_\mu \mu\varphi_\mu = n$.

Let $A \overset{\text{def}}{=} A_\Phi \overset{\text{def}}{=} cn^{-1+a} = cn^a/n$ and $B \overset{\text{def}}{=} 4n^{-1+0.6a} = 4n^{0.6a}/n$. Then, let $\gamma \overset{\text{def}}{=} n^{-1.5}$ and let $\mathcal{X} \overset{\text{def}}{=} \{\gamma, 2^2\gamma, 3^2\gamma, \ldots, A + B/2\} = \{x_\ell = \ell^2 n^{-1.5} : \ell = 1, 2, \ldots, |\mathcal{X}|\}$, where clearly, $|\mathcal{X}| = \sqrt{(A+B/2)n^{1.5}} = \mathcal{O}(n^{0.25+0.5a})$.

A linear program corresponding to $\overline{\varphi}$ is defined as follows.

Definition 46. (Also [67, Definition 18].) Consider a linear program consisting of variables $v_x \ge 0$ for all $x \in \mathcal{X}$ satisfying the following three constraints:

C1. $\sum_{x \in \mathcal{X}: A \le x \le A+B/2} x v_x \le 16n^{-0.4a}$.

C2. $\sum_{x \in \mathcal{X}} x v_x + \sum_{\mu > (A+B)n} \frac{\mu}{n}\varphi_\mu = 1$.

C3. For all $\mu \in \{1, 2, \ldots, (A+B/4)n\}$,

$$\sum_{x \in \mathcal{X}} v_x \text{poi}(nx, \mu) \in [\varphi_\mu - 4n^{0.99+a}, \varphi_\mu + 4n^{0.99+a}].$$

Any solution v to the linear program is extended to obtain a corresponding histogram h^v as follows. We create a generalized histogram h' whose lower part consists of v and upper part consists of empirical frequencies (as indicated in C2). However, since the heights of a proper histogram can only be integers, the v_x's are rounded down to their nearest integers, and the entire support is scaled suitably by $(1+\epsilon)$ so that the mass adds up to 1.

Definition 47. (Also [67, Definition 19].) Any solution solution v of the linear program is extended to a proper histogram as follows.

S1. Set $h'(x) = 0$ and $h^v(x) = 0$ for all x.

S2. For all $x \in \mathcal{X}$, set $h'(x) = v_x$, and for all $\mu \geq (A+B)n$, set $h'(\mu/n) = \varphi_\mu$.

S3. For all $x \in (0,1]$ such that $h'(x) \neq 0$, set $h^v((1+\epsilon)x) = \lfloor h'(x) \rfloor$, where $\epsilon = \frac{\sum_{x\in\mathcal{X}} x(v_x - \lfloor v_x \rfloor)}{1-\sum_{x\in\mathcal{X}} x(v_x - \lfloor v_x \rfloor)}$. □

Note that h' is the generalized histogram such that $h'(x) = v_x$ for $x \in \mathcal{X}$ and $h'(\mu/k) = \varphi_\mu$ for $\mu \geq (A+B)n$. In Step S3, the choice of ϵ is such that $\sum_y y h^v(y) = 1$. The estimator is thus defined as follows.

Definition 48. (Also [67, Algorithm].) Given $\overline{X}$ whose profile is $\varphi(\overline{X}) = \overline{\varphi}$, the estimator Q^{VV} outputs a distribution multiset, *i.e.*, equivalently its histogram, by the following steps.

T1 Construct the linear program of Definition 46 corresponding to $\overline{\varphi}$.

T2 Find a solution v to the linear program. If no solution exists, output ***fail***.

T3 Output the histogram h^v corresponding to v, as given by Definition 47.

The correctness of this estimator is shown by the following result.

Theorem 49. *(Also [67, Theorem 3].) For any constant* $\delta \in (0,1]$, *and sufficiently large* n *(as a function of* δ*), in particular* n *at least* $\exp(1/\delta)$, *given* n i.i.d. *samples from any distribution* h *of support size* $k \leq \delta n \log n$, *with probability*

$\geq 1 - \exp(-n^{0.94})$*, the VV estimator outputs a histogram* h^v *such that the relative earthmover distance between* h *and* h^v *is*

$$R(h, h^v) \leq 70000\sqrt{\delta} \cdot \max\{1, |\log(\delta)|\}.$$

□

Corollary 50. *For all sufficiently large* n*,* $\epsilon > 0$*, and all distributions* P *of support size* $k = \mathcal{O}(\epsilon^{2.1} n \log(n))$*, if* $\overline{X} \sim P^n$*, then* $\Pr(R(P, Q^{\mathrm{VV}}_{\overline{X}}) > \epsilon) \leq e^{-n^{0.91}}$*. Since* $R(P, P') \geq \frac{1}{2}|P - P'|_1$*, it follows that* $\Pr(|P - Q^{\mathrm{VV}}_{\overline{X}}|_1 > 2\epsilon) \leq e^{-n^{0.91}}$*.* □

The detailed proof of Theorem 49 can be found in [67]. While an error probability of $e^{-n^{0.03}}$ is shown in [67], it can be shown similarly that the modified estimator Q^{VV} has error probability $e^{-n^{0.94}}$ as stated here. Most steps in the proof are straightforward to extend so as to ensure that the error probabilities at various steps is at most $e^{-n^{0.98}}$. But we do mention that [67, Lemma 25] and [67, Proposition 17] require additional care. Specifically, in [67, Lemma 25], it involves the step that shows the relative earthmover distance contributed by low probability symbols that appear high number of times (and hence estimated by empirical frequencies) is small, $o_n(1)$, with high probability $\geq 1 - e^{-n^{0.98}}$. In [67, Proposition 17], it involves showing that when high probabilities are approximated by empirical estimates, the earthmover distance incurred is $o_n(1)$ w.p. $\geq 1 - e^{-n^{0.98}}$. The former requires dividing the range of observed counts and latter requires dividing the range of probabilities, followed by application of Chernoff bounds and union bound over each of these parts.

We also mention that the most difficult parts of the proof, [67, Lemma 24] and [67, Appendix C.2], that involve showing that any two solutions to the linear program of the estimator are close in earthmover distance, are straightforward to extend (by tweaking various constants), mainly because they are purely analytical and do not involve any probability calculations. It essentially captures the earlier motivation that $\mathrm{E}_Q[\overline{\varphi}] \approx \overline{\varphi} \approx \mathrm{E}_P[\overline{\varphi}]$ implies $\mathcal{M}(P) \approx \mathcal{M}(Q)$, *i.e.,* $R(P, Q) \leq \epsilon$. On that note, we make the following observation.

Lemma 51. *For two distributions P and P', the expected length-n profiles are same iff their distributions on length n profiles are the same,* i.e.,

$$\mathrm{E}_P[\overline{\varphi}] = \mathrm{E}_{P'}[\overline{\varphi}] \;\Leftrightarrow\; P(\Phi^n) = P'(\Phi^n).$$

Proof Sketch. We show the forward direction. It is easy to see by induction that if $\mathrm{E}_P[\varphi_\mu] = \mathrm{E}_{P'}[\varphi_\mu]$, *i.e.,*

$$\sum_{i=1}^{k} \binom{n}{\mu} p_i^\mu (1-p_i)^{n-\mu} = \sum_{i=1}^{k} \binom{n}{\mu} p_i'^\mu (1-p_i')^{n-\mu},$$

for $\mu = 1, 2, \ldots, n$, then the power sums $S_\mu(P) \overset{\text{def}}{=} \sum_i p_i^\mu = \sum_i p_i'^\mu \overset{\text{def}}{=} S_\mu(P')$ are the same for $\mu = 1, 2, \ldots, n$. It is a well known fact in algebra, *e.g.,* [17], that power sums form a basis for symmetric polynomials of (total) degree at most n. Since profile probabilities $P(\overline{\varphi}), P'(\overline{\varphi})$, are monomial symmetric polynomials of degree n as we saw in Section 2.4, it implies that $P(\Phi^n) = P'(\Phi^n)$. The other direction is easy to show by probability or algebraic arguments. □

For a simple example, the distributions $P = (\frac{1}{2}, \frac{1}{2})$ and $P' = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$ are indistinguishable on profiles of length-2 sequences since $\sum_i p_i = \sum_i p_i' = 1$ and $\sum_i p_i^2 = \sum_i p_i'^2 = \frac{1}{2}$. (Distributions of support size $k \le 3$ can be visualized on the simplex $p_1 + p_2 + p_3 = 1$. All distributions on the circles obtained by intersecting spheres $p_1^2 + p_2^2 + p_3^2 = c$ with the simplex (plane) are indistinguishable on length-2 sequences. Larger alphabet distributions and profiles of longer sequences, *i.e.,* higher moments are harder to visualize.) In general, it is useful to find distributions P and P' of support size k, which are very different, say their property values $\pi(P)$ and $\pi(P')$ are very different for some property π, but $|P(\Phi^n) - P'(\Phi^n)|$ is small. In various works on testing and estimating properties of distributions, *e.g.,* [7, 55, 70, 68], this is the technique used for showing lower bounds on sample complexity n in terms of alphabet size k.

5.2 Other approaches to distribution estimation

Calculation and approximation of profile maximum likelihood appears to be difficult in general, *e.g.,* see [50]. Many results, *i.e.,* support sizes and number of distinct probabilities of the PML distribution, are known for special kinds of profiles, *e.g.,* see [50, 45, 5, 6, 3]. Exact calculation of PML using such bounds for profiles of smaller lengths using algebraic elimination methods, namely Groebner bases and resultants, is discussed briefly in Section 5.3.

In this section, we consider alternative approaches to distribution estimation, that are motivated by both computational efficiency and other criteria for measuring the quality of distribution estimates. We consider a distinguishability based criteria for evaluating the quality of estimators. It is natural to want an estimator Q such that length-n *i.i.d.* sequences generated from it are indistinguishable from those generated by the actual distribution P. Following the discussion in Section 2.5, if we can find an estimator Q such that

$$|P(\Phi^n) - Q(\Phi^n)| \stackrel{\text{def}}{=} \sum_{\overline{\varphi}\in\Phi^n} |P(\overline{\varphi}) - Q(\overline{\varphi})|,$$

is small and close to 0, or equivalently the affinity,

$$|P(\Phi^n) \wedge Q(\Phi^n)| \stackrel{\text{def}}{=} \sum_{\overline{\varphi}\in\Phi^n} \min\{P(\overline{\varphi}), Q(\overline{\varphi})\},$$

is large and close to 1, then P and Q cannot be distinguished on Φ^n by any test.

However, it is easy to see that we may not hope for such estimators, even for distributions of small support. For example, if $P = (\frac{3}{4}, \frac{1}{4})$ and n is large, the empirical estimator outputs $Q = (\frac{3}{4}, \frac{1}{4}) + \Omega(\frac{1}{\sqrt{n}})$ with probability $\frac{1}{10}$. But length-n sequences from P and such Q can be distinguished with error probability $\leq \frac{1}{4}$, *i.e.,* affinity $|P(\Phi^n)\wedge Q(\Phi^n)| \leq \frac{1}{2}$. This is easier to see in the Poisson model where we are given poi(n) samples using the fact that for large λ, $|\text{poi}(\lambda) - \text{poi}(\lambda + \Omega(\sqrt{\lambda}))| \geq 1$. However, empirical estimator is essentially the best estimator in this case. Equivalently, no estimator Q can have affinity close to 1 for both $P = (\frac{3}{4}, \frac{1}{4})$ and $P' = (\frac{3}{4}, \frac{1}{4}) + \Omega(\frac{1}{\sqrt{n}})$. We can similarly construct examples of distributions P that have many large probabilities spaced far apart, and empirical estimator Q is essentially the best one can do, yet $|P(\Phi^n) \wedge Q(\Phi^n)|$ is arbitrarily small.

We note that if $|P(\Phi^n) \wedge Q(\Phi^n)|$ is small, say $\leq \frac{1}{2}$, it does not necessarily imply that Q is a bad estimate of P. This is because $|P(\Phi^n) \wedge Q(\Phi^n)|$ is error probability of distinguishing P and Q in a simple hypothesis testing problem. In general, we would like to compare against closeness tests that do not know about P and Q. In such a closeness testing problem, one is given two sequences $\overline{X} \sim P^n$ and $\overline{X}' \sim P'^n$ and asked whether $\mathcal{M}(P)$,$\mathcal{M}(P')$ are same or different. However, it is hard to analyze such closeness tests.

We also observe that if we can find an estimator Q such that $|P(\Phi^n) \wedge Q(\Phi^n)| \geq e^{-10\sqrt{n}}$, it still implies $R(P,Q) \leq 3\epsilon$ when $k = \mathcal{O}(\epsilon^{2.1} n \log(n))$. Otherwise, if $R(P,Q) > 3\epsilon$, then the test $R(P, Q^{\text{VV}}_{\varphi(\overline{X})}) \underset{Q}{\overset{P}{\lessgtr}} R(Q, Q^{\text{VV}}_{\varphi(\overline{X})})$ can distinguish between P and Q with error probability $e^{-n^{0.9}}$, contradicting $|P(\Phi^n) \wedge Q(\Phi^n)| = e^{-10\sqrt{n}}$. We currently do not have any computationally efficient estimators that have such error guarantees. Unsurprisingly, the PML estimator offers such guarantees by the following simple reasoning. For any P, if $\overline{X} \sim P^n$, then $P(P(\varphi(\overline{X})) \leq e^{-4\sqrt{n}}) \leq e^{-4\sqrt{n}}|\Phi^n| \leq e^{-\sqrt{n}}$. Thus, with probability $\geq 1 - e^{-\sqrt{n}}$, the generated $\overline{X}$ is such that $Q^{\text{PML}}(\varphi(\overline{X})) = \hat{P}(\varphi(\overline{X})) \geq P(\varphi(\overline{X})) \geq e^{-4\sqrt{n}}$, implying that $|P(\Phi^n) \wedge Q^{\text{PML}}(\Phi^n)| \geq \min\{P(\varphi(\overline{X})), Q^{\text{PML}}(\varphi(\overline{X}))\} \geq e^{-4\sqrt{n}}$.

Although it is difficult to construct an estimator that guarantees $|P(\Phi^n) \wedge Q(\Phi^n)| \geq e^{-10\sqrt{n}}$, we build upon Q^{VV} and utilize several concentration properties of profiles of *i.i.d.* sequences to construct an estimator Q that is harder to distinguish from P compared to Q^{VV}. We note that for many distributions P of alphabet size $k = o(n)$, Q^{VV} can be easily distinguished from P, since it uses empirical estimates for high probability symbols. This is shown in the following example.

Example 52. or sufficiently large n, let $P = U[n^{0.9}]$ be the uniform distribution on $k = n^{0.9}$ symbols. If $\overline{X} \sim P^{\text{poi}(n)}$, $\mathrm{E}_P[\varphi_\mu] = n^{0.9} \cdot \text{poi}(n^{0.1}, \mu)$ for $\mu \in \{1, 2, \ldots\}$. Since the prevalences concentrate around their mean, with high probability, the empirical distribution Q, and hence Q^{VV}, has roughly $\varphi_\mu \approx \mathrm{E}_P[\varphi_\mu]$ number of $\frac{\mu}{n}$.

We then have

$$\begin{aligned}
\mathrm{E}_Q[\varphi_{n^{0.1}}] &\approx \sum_{\mu=1}^{\infty} \mathrm{E}_P[\varphi_\mu] \cdot \mathrm{poi}\big(n \cdot \frac{\mu}{n}, n^{0.1}\big) \\
&= n^{0.9} \sum_{\mu=1}^{\infty} \mathrm{poi}(n^{0.1}, \mu) \cdot \mathrm{poi}\big(\mu, n^{0.1}\big) \\
&\approx \frac{1}{2} n^{0.9} \mathrm{poi}(n^{0.1}, n^{0.1}) \\
&= \frac{1}{2} \mathrm{E}_P[\varphi_{n^{0.1}}],
\end{aligned}$$

using the fact that for large λ, $\sum_{\mu=0}^{\infty} \mathrm{poi}(\lambda, \mu)\mathrm{poi}(\mu, \lambda) \approx \frac{1}{2}\mathrm{poi}(\lambda, \lambda)$. As $\mathrm{E}_P[\varphi_{n^{0.1}}] = n^{0.9} \cdot \mathrm{poi}(n^{0.1}, n^{0.1}) \approx \frac{n^{0.9}}{\sqrt{2\pi n^{0.1}}} \geq n^{0.8}$. Hence, we conclude using Observation 44 that P and Q can be distinguished by the test $\varphi_{n^{0.1}} \underset{Q}{\overset{P}{\lessgtr}} \frac{3}{4}\mathrm{E}_P[\varphi_{n^{0.1}}]$ with error probability at most $e^{-n^{0.6}}$. This also implies $|P(\Phi^n) \wedge Q(\Phi^n)| \leq e^{-n^{0.6}}$. □

In the above example, we see that if an estimator Q uses empirical estimates for high probabilities, $\mathrm{E}_Q[\varphi_\mu]$ can be far from $\mathrm{E}_P[\varphi_\mu]$ and thus, P and Q are easily distinguishable. We therefore attempt to improve upon empirical estimation of high probabilities by making $\mathrm{E}_Q[\varphi_\mu] \approx \varphi_\mu$, even for large μ. To do this, we observe the following concentration properties of profiles that are similar to 44.

Observation 53. *Let $b \in (0, 1)$ be a small constant. For any P, let $\overline{X} \sim P^{\mathrm{poi}(n)}$ and $\varphi(X) = (\varphi_1, \varphi_2, \ldots)$. If $\mathrm{E}[\varphi_\mu] \geq n^b$, then*

$$\Pr\big(|\varphi_\mu - \mathrm{E}[\varphi_\mu]| \geq (\mathrm{E}[\varphi_\mu])^{0.6}\big) \leq 2\exp(-n^{0.2b}/4),$$

and if $\mathrm{E}[\varphi_\mu] < n^b$, then

$$\Pr\big(\varphi_\mu \geq 2n^b\big) \leq \exp(-n^b/4).$$

Proof. Using Chernoff bounds on $\varphi_\mu = \sum_{i=1}^{k} \mathbb{1}_{[\mu(a_i)=\mu]}$, which is a sum of independent 0-1 random variables (in the Poisson model). □

Observation 54. *Let $b \in (0, 1)$ be a small constant. For any P, let $\overline{X} \sim P^{\mathrm{poi}(n)}$. For all $a \in \mathcal{A}$, if $nP(a) > n^b$, then*

$$\Pr\big(|\mu(a) - nP(a)| \geq (nP(a))^{0.6}\big) \leq 2\exp(-n^{0.2b}/4),$$

and if $nP(a) < n^b$*, then*

$$\Pr\left(\mu(a) \geq 2n^b\right) \leq \exp(-n^b/4).$$

Proof. Using Poisson tail bounds on $\mu(a) \sim \text{poi}(nP(a))$. □

Thus, we modify Q^{VV} to solve for the following set of constraints, whose solution we know exists with high probability by above observations.

Definition 55. Given a profile $\overline{\varphi} = (\varphi_1, \ldots, \varphi_n) = \{\mu_1, \mu_2, \ldots, \mu_m\}$ such that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m > 0$. Let b_1, b_2 be small constants. Let $\ell = |\{\mu_i \geq n^{b_1}\}|$ be the number of multiplicities larger than n^{b_1}. Note that $\mu_1, \ldots, \mu_\ell$ are these multiplicities. Let $Q = \{q_1, q_2, \ldots, q_k\}$ be a distribution that satisfies the following constraints:

C1 For $i = 1, \ldots, \ell$, $|nq_i - \mu_i| \leq \mu_i^{0.6}$.

C2 $k \leq n^2$.

C3 For $i = \ell, \ell+1, \ldots, k$, $nq_i \leq 2n^{b_1}$.

C4 $\sum_i q_i = 1$.

C5 For $\mu = 1, 2, \ldots, n$, if $\varphi_\mu \geq n^{b_2}$, $|\mathrm{E}_Q[\varphi_\mu] - \varphi_\mu| \leq \varphi_\mu^{0.6}$ and if $\varphi_\mu < n^{b_2}$, $\mathrm{E}_Q[\varphi_\mu] < 2n^{b_2}$. □

We can find such a Q by linear programming similar to Q^{VV}. For low probabilities $\{q_\ell, \ldots, q_k\}$, our program is similar. For large probabilities, $\{q_1, \ldots, q_\ell\}$, satisfying C1 and C5 simultaneously is difficult. One possible way of solving for the large q_i, $i \in \{1, \ldots, \ell\}$ is as follows. Let each such q_i be allowed to take values only in the set $\{q_{i,j} = \frac{\mu_i - \mu_i^{0.6}}{n} + \frac{j}{n^2} : j = 0, 1, \ldots, 2n\mu_i^{0.6}\}$, *i.e.,* one among a discrete set of values in the range $[\frac{\mu_i - \mu_i^{0.6}}{n}, \frac{\mu_i + \mu_i^{0.6}}{n}]$. This can be translated into an integer program by creating 0-1 variables $x_{i,j}$, *i.e.,* by introducing integer variables $x_{i,j}$ and constraints $x_{i,j} \geq 0$ and $\sum_j x_{i,j} = 1$ corresponding to C1. Then C5 translates to the linear constraints

$$\sum_{i=1}^{\ell} \sum_{j=0}^{2n\mu_i^{0.6}} x_{i,j}\mathrm{poi}(nq_{i,j},\mu) \in [\varphi_\mu - \varphi_\mu^{0.6}, \varphi_\mu - \varphi_\mu^{0.6}] \quad \text{or} \quad \leq 2n^{b_2}$$

depending on whether $\varphi_\mu \gtrless n^{b_2}$. We can either solve these integer linear constraints or solve for the relaxation that $x_{i,j}$ are real, instead of integers. If such a solution is found, then a natural candidate for q_i is $q_i = \sum_j x_{i,j}q_{i,j}$. However, such an estimate may not work since $\mathrm{poi}(nq_i,\mu) = \mathrm{poi}\big(n\sum_j x_{i,j}q_{i,j},\mu\big)$ may not be a good approximation of $\sum_j x_{i,j}\mathrm{poi}(nq_{i,j},\mu)$ that was solved for in the constraints. (This is because the derivative of Poisson function $\mathrm{poi}(x,\mu)$ is too sharp to allow for good linear approximation within $x \pm x^{0.6}$.)

An alternative is to refine upon the estimates q_i in multiple rounds, where we start with $q_i = \frac{\mu_i}{n}$ in round 1, and in round $r \in \{1,2,\ldots\}$, we exponentially narrow down our range of $q_{i,j}$ to $[\frac{nq_i}{n} - \frac{1}{2^{r-1}}\frac{(nq_i)^{0.6}}{n}, \frac{nq_i}{n} + \frac{1}{2^{r-1}}\frac{(nq_i)^{0.6}}{n}]$. We terminate the process in some round if the q_i's obtained are a solution to the constraints, or the constraints have no solution, or if the number of rounds is $\mathcal{O}(\log n)$. For simplicity, instead of many $x_{i,j}$ and $q_{i,j}$ for each i, we may use two variables x_i^- and x_i^+ corresponding to $q_i^\mp = \frac{nq_i}{n} \mp \frac{1}{2^{r-1}}\frac{(nq_i)^{0.6}}{n}$.

5.3 Exact calculation of PML by algebraic elimination

In this section, we consider exact calculation of PML using elimination methods from algebra. Recall from Section 2.4 that pattern and profile probabilities are symmetric polynomials in the probabilities $\mathcal{M}(P) = \{p_1,\ldots,p_k\}$. Hence, given an upper bound on the support size of the PML distribution, we maximize the pattern probability by solving the system of multivariate polynomial equations obtained by differentiating the pattern probability with respect to each variable. As earlier, let

$$\mathcal{P}_k \overset{\text{def}}{=} \{(p_1,p_2,\ldots,p_k) : p_i \geq 0, \sum_{i=1}^{k} p_i = 1\}$$

be the class of distributions whose support size is at most k. The Kuhn-Tucker conditions imply that if a distribution $P \in \mathcal{P}_k$ is a local maximum of $P(\overline{\psi})$, then for all $i, j \in [k]$ such that $p_i, p_j > 0$,

$$\frac{\partial P(\overline{\psi})}{\partial p_i} = \frac{\partial P(\overline{\psi})}{\partial p_j}.$$

To find $\hat{P}_{k,\overline{\psi}} \overset{\text{def}}{=} \arg\max_{P \in \mathcal{P}_k} P(\overline{\psi})$, we can therefore solve the system of equations

$$\frac{\partial P(\overline{\psi})}{\partial p_i} = \frac{\partial P(\overline{\psi})}{\partial p_{i+1}}, \quad i = 1, 2, \ldots, \kappa - 1 \tag{I}$$

$$p_1 + p_2 + \cdots + p_\kappa = 1, \tag{II}$$

for each $\kappa \in \{m, m+1, \ldots, k\}$, and among all solutions, find the one maximizing $P(\overline{\psi})$. Note that since the Kuhn-Tucker equalities hold only for the nonzero probabilities, it is not sufficient to consider solutions to system of equations (I) and (II) for just $\kappa = k$.

As an example of this basic method and the need to solve the equations for all $\kappa \leq k$, consider $\hat{P}_{3,112}$. For $\kappa = 2$, the equations yield

$$\begin{aligned} &p_1^2 = p_2^2, \\ &p_1 + p_2 = 1, \end{aligned}$$

whose unique solution is $P = (p_1, p_2) = (\frac{1}{2}, \frac{1}{2})$ with $P(112) = \frac{1}{4}$. For $\kappa = 3$, Equations (I) and (II), yield

$$\begin{aligned} &p_1(2 - 3p_1) = p_2(2 - 3p_2) = p_3(2 - 3p_3), \\ &p_1 + p_2 + p_3 = 1, \end{aligned}$$

whose only solution is $P' = (p_1, p_2, p_3) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Furthermore, $P'(112) = \frac{2}{9} < \frac{1}{4} = P(112)$, hence $\hat{P}_{3,112} = (\frac{1}{2}, \frac{1}{2})$.

While such simple manipulations work for small patterns, for longer patterns, we need a systematic approach for solving the set of polynomial equations obtained. The natural approach for solving a system of polynomial equations is to generalize the Gaussian-elimination method for linear equations to address polynomial equations. There are two well-known approaches for doing that.

The first uses Buchberger's algorithm and its variations that yield a *Groebner basis* for the original polynomials. However, the degrees of the resulting polynomials, and hence also the computation time of these algorithms may in general be doubly exponential in the number of variables,*e.g.,* see [17].

The second approach uses resultants, *e.g.,* [16, 17]. While it too may require doubly-exponential time, in our experiments it has performed more efficiently, and we describe it here.

The *resultant* of a degree-u polynomial $f = f_0 x^u + f_1 x^{u-1} + \cdots + f_u$ and a degree-v polynomial $g = g_0 x^v + g_1 x^{v-1} + \cdots + g_v$ is the determinant of a corresponding $(u+v) \times (u+v)$ *Sylvester matrix*,

$$\operatorname{Res}(f,g,x) \overset{\text{def}}{=} \begin{vmatrix} f_0 & & & & g_0 & & & \\ f_1 & f_0 & & & g_1 & g_0 & & \\ \vdots & f_1 & \ddots & f_0 & \vdots & g_1 & \ddots & g_0 \\ f_u & \vdots & \ddots & f_1 & g_v & \vdots & \ddots & g_1 \\ & f_u & & \vdots & & g_v & & \vdots \\ & & \ddots & f_u & & & \ddots & g_v \end{vmatrix},$$

where v columns correspond to f, u columns to g, and blank spaces are zeros.

If f and g are multivariate polynomials with x as one of the variables, then viewing f and g as polynomials in x whose coefficients are polynomials in the other variables, $\operatorname{Res}(f,g,x)$ is a polynomial in the remaining variables. The important property of resultants that makes them useful for elimination is that $\operatorname{Res}(f,g,x) = a \cdot f + b \cdot g$ for two polynomials a and b. Hence solving the equations $f = g = 0$ is equivalent to solving the system $f = \operatorname{Res}(f,g) = 0$.

To eliminate variables $p_2, \ldots, p_{\kappa-1}$ from Equations (I), we use resultants to eliminate p_2 from the $\kappa - 1$ equations in (I) to obtain $\kappa - 2$ equations in $p_1, p_3, \ldots, p_\kappa$ from which we eliminate p_3 and obtain $\kappa - 3$ equations. We proceed similarly to eliminate $p_4, \ldots, p_{\kappa-1}$, until we are left with a single homogeneous equation in p_1 and p_κ. We use this to solve for $\frac{p_\kappa}{p_1}$ and by backsubstitution, obtain $(\frac{p_{\kappa-1}}{p_1}, \frac{p_{\kappa-2}}{p_1}, \ldots, \frac{p_2}{p_1})$. Finally, using Equation (II), we obtain all the probabilities $(p_1, p_2, \ldots, p_\kappa)$.

While the resultant of two polynomials can be obtained by explicitly computing the determinant of their Sylvester matrix or using other well known determinantal formulae for resultants, a brute-force computation is challenging. Other computationally efficient methods for computing resultants that use interpolation techniques are discussed in [40, 16]. Nevertheless, it is easy to see that after eliminating $p_2, \ldots, p_{\kappa-1}$, the degree of the final polynomial in p_1 and p_κ can be $O(n^{2^\kappa})$. This makes the resultant calculations intensive for even small values of n and κ.

The number of calculations is smaller when considering distributions with at most Δ distinct probabilities. For such distributions, it suffices to consider for each $d \in 1, 2, \ldots, \Delta$, partitions of $\{p_1, p_2, \ldots, p_\kappa\}$ into d parts where within each part all probabilities are equal, and perform the elimination with d variables.

When evaluating resultants of two polynomials, we remove their common factors, otherwise the resultant is zero. While we do not discuss mixed distributions that have discrete probabilities as well as a continuous part, the method can be easily extended to this case by adding an additional variable q for the probability of the continuous part. The complete method is summarized in Algorithm 1, which computes the PML distribution $\hat{P}_{k,\Delta,\overline{\psi}}$, given as input a pattern $\overline{\psi}$ and bounds k and Δ on alphabet size and number of distinct probabilities.

We implemented Algorithm 1 in Mathematica. Due to computational limitations, the program can be used to compute $\hat{P}_{k,\Delta,\overline{\psi}}$ for patterns of length ≤ 14 with $k \leq 17$ and and $\Delta \leq 4$. While we do not have good general upper bounds on $\hat{k}$ and $\hat{\Delta}$ two plausible assumptions are:

A1 $\hat{\Delta}$ is at most the number of distinct multiplicities in the pattern.

A2 $\hat{P}_k(\overline{\psi})$ is strictly increasing for $m \leq k \leq \hat{k}$.

One possible justification for A1 is that each of the probabilities may correspond to an observed symbol and then symbols whose multiplicities are equal would be assigned equal probability estimates. For A2, it may be plausible that if P_1 and P_2 are two distributions whose alphabet sizes are k and $k+2$, and $P_1(\overline{\psi}) > P_2(\overline{\psi})$, then there may exist a distribution P_3 whose support size is $k+1$ such that $P_1(\overline{\psi}) < P_3(\overline{\psi}) \leq P_2(\overline{\psi})$.

Under these assumptions, given a pattern $\overline{\psi}$, we use Algorithm 1 with Δ as the number of distinct multiplicities in $\overline{\psi}$ for $k = m, m+1, \ldots$ until $\hat{P}_{k,\Delta,\overline{\psi}}(\overline{\psi}) = \hat{P}_{k+1,\Delta,\overline{\psi}}(\overline{\psi})$.

For example, for the canonical pattern $\overline{\psi} = 1^5 2^2 3^2 45$ of *abracadabra*, the multiplicities of the symbols are $(5, 2, 2, 1, 1)$ and there are 3 distinct multiplicities 5, 2 and 1. Assumption A1 implies that the number of distinct probabilities is at most 3. Since $m = 5$, we run Algorithm 1 with $\Delta = 3$ and $k = 5, 6, 7, \ldots$, and observe that $\hat{P}_{5,3,1^5 2^2 3^2 45} = 3.241..\times 10^{-6}$, $\hat{P}_{6,3,1^5 2^2 3^2 45} = \hat{P}_{7,3,1^5 2^2 3^2 45} = 4.073..\times 10^{-6}$. Hence we stop and output $\hat{P}_{6,3,1^5 2^2 3^2 45} = \left\{\frac{\alpha}{5+\alpha}, (\frac{1}{5+\alpha})^5\right\} = \{0.4429.., 0.1114..^5\}$, where $\alpha = 3.976..$ is a root of $6x^4 - 19x^3 - 19x^2 - x - 1 = 0$.

Under these assumptions, we computed the PML of all patterns of length ≤ 14. For space considerations, Table 5.1 shows the PML only of patterns of length ≤ 10. Furthermore, the PML of all binary, ternary, skewed, and quasi-uniform patterns have been determined before, and the table shows the remaining patterns. The PML distribution $\hat{P}_{\overline{\psi}}$ is represented in the form $\{\tilde{p}_1^{k_1}, \ldots, \tilde{p}_d^{k_d}\}, q$ indicating that for $i = 1, \ldots, d$ it consists of k_i symbols whose probability is $\tilde{p}_i$, and that the continuous part is $q = 1 - \sum_{i=1}^{d} k_i \tilde{p}_i$, shown only when nonzero. Note that all numbers are algebraic, and are truncated to a few significant digits.

Acknowledgement

Chapter 5 is adapted from Jayadev Acharya, Hirakendu Das, Alon Orlitsky, Shengjun Pan, "Algebraic computation of pattern maximum likelihood", *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 2011; and Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, Ananda Theertha Suresh, "Competitive estimation of discrete probability distributions", In Preparation, 2012. The dissertation author was the primary researcher and author of this paper.

Table 5.1: PML of patterns of length ≤ 10, under assumptions.

n	$\overline{\psi}$	$\hat{P}_{\overline{\psi}}$	$\hat{P}_{\overline{\psi}}(\overline{\psi})$
8	$1^4 2^2 34$	$\{0.462.., 0.134..^4\}$	4.08×10^{-4}
	$1^3 2^3 34$	$\{0.25^4\}$	3.66×10^{-4}
	$1^3 2^2 345$	$\{0.15625^6, 0.0625\}$	4.38×10^{-4}
9	$1^5 2^2 34$	$\{0.553.., 0.0893..^5\}$	1.99×10^{-4}
	$1^4 2^3 34$	$\{0.389..^2\}, 0.222..$	1.33×10^{-4}
	$1^4 2^2 3^2 4$	$\{0.25^4\}$	9.16×10^{-5}
	$1^4 2^2 345$	$\{0.433.., 0.0708..^8\}$	1.10×10^{-4}
	$1^3 2^3 3^2 4$	$\{0.25^4\}$	9.16×10^{-5}
	$1^3 2^3 345$	$\{0.333..^2\}, 0.333..$	1.02×10^{-4}
	$1^3 2^2 3456$	$\{0.138..^6\}, 0.171..$	1.57×10^{-4}
10	$1^6 2^2 34$	$\{0.599.., 0.0800..^5\}$	1.15×10^{-4}
	$1^5 2^3 34$	$\{0.4^2\}, 0.2$	5.24×10^{-5}
	$1^5 2^2 3^2 4$	$\{0.461.., 0.180..^3\}$	2.67×10^{-5}
	$1^5 2^2 345$	$\{0.499.., 0.0626..^5\}$	5.02×10^{-5}
	$1^4 2^4 34$	$\{0.4^2\}, 0.2$	5.24×10^{-5}
	$1^4 2^3 3^2 4$	$\{0.25^4\}$	2.29×10^{-5}
	$1^4 2^3 345$	$\{0.35^2\}, 0.3$	3.47×10^{-5}
	$1^4 2^2 3^2 45$	$\{0.304.., 0.139..^5\}$	1.23×10^{-5}
	$1^4 2^2 3456$	$\{0.3^2\}, 0.4$	3.73×10^{-5}
	$1^3 2^3 3^3 4$	$\{0.25^4\}$	2.29×10^{-5}
	$1^3 2^3 3^2 45$	$\{0.2^5\}$	1.23×10^{-5}
	$1^3 2^3 3456$	$\{0.3^2\}, 0.4$	3.73×10^{-5}
	$1^3 2^2 34567$	$\{0.157..^4\}, 0.371..$	7.16×10^{-5}

Algorithm 1 Computation of $\hat{P}_{k,\Delta,\overline{\psi}}$ using resultants

```
Initialize solution set $\mathcal{S} = \{\}$
for $\kappa = m$ to $k$ and $d = 1$ to $\max\{\Delta, \kappa\}$ do
  for unordered partitions $(\mathcal{K}_1, \ldots, \mathcal{K}_d)$ of $[\kappa]$ do
    for $i = 1$ to $d - 1$ do
      $g_{1,i} := \frac{\partial P(\overline{\psi})}{\partial p_\imath} - \frac{\partial P(\overline{\psi})}{\partial p_\jmath}$ for some $\imath \in \mathcal{K}_i$ and $\jmath \in \mathcal{K}_{i+1}$
      Set $p_\imath = z_j$ for all $\imath \in \mathcal{K}_j$ for all $j$
    end for
    $\mathcal{G}_1 := \{g_{1,1}, g_{1,2}, \ldots, g_{1,d-1}\}$
    // Solve the system of equations $\mathcal{G}_1 = 0$
    for $j = 2$ to $d - 1$ do
      for $i = 1$ to $d - j$ do
        // Eliminate $z_j$ by taking resultants
        $g_{j,i} = \mathrm{Res}(g_{j-1,i}, g_{j-1,i+1}, z_j)$
        Remove any monomial factors and trivial factors $(z_\imath - z_\jmath)$ of $g_{j,i}$
        Return error for current $k$ and $d$ if $g_{j,i}$ is a constant polynomial
      end for
    end for
    Set $z_1 = 1$ and solve the triangular system of equations $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_{d-1}\}$
    for variables $z_2, z_3, \ldots, z_d$ by backsubstitution
    For solutions $(z_1, z_2, \ldots, z_d)$ such that all $z_i$ are real and positive, find $Z = \sum_{i=1}^{d} z_i |\mathcal{K}_i|$, find $(p_1, p_2, \ldots, p_\kappa)$ such that $p_\imath = z_j / Z$ where $\imath \in \mathcal{K}_j$ and $S = S \cup \{sort(p_1, p_2, \ldots, p_\kappa)\}$
  end for
end for
Output $\arg\max_{P \in \mathcal{S}} P(\overline{\psi})$
```

Chapter 6

Bernoulli and Poisson Multiset Estimation

Many problems of probability estimation involve parametric families of distributions. Often, one is interested in estimating the *multiset* of parameters of the distribution, given samples generated by it. Some of these problems are related to distribution multiset estimation problem considered in Chapter 5. We consider two such problems here, which we call the Bernoulli multiset estimation and the Poisson multiset estimation.

Both problems are motivated by applications of load estimation by service providers, say a website or telephone company that wants to estimate the usage pattern of its users. In the Bernoulli multiset estimation problem, we want to find the multiset of success probabilities of multiple independent Bernoulli processes. Each process corresponds to a user and takes values 0 or 1 at different times, indicating whether the user is active or not at those times. In the Poisson multiset case, we want to estimate the means of a collection of independent Poisson distributions given a sample from each of them. The activities of different users are modeled as independent Poisson processes and observed over a fixed time period.

In the next two sections, we consider the problems in detail. We show that the PML estimator for these problems is competitive with any other estimator whose error probability is exponentially or near exponentially small. Furthermore, their close relationship with the distribution estimation problem can be used to show constructively the existence of estimators with such small error probabilities. Thus, good distribution estimators can be used to construct good Bernoulli and Poisson multiset estimators as well. We first consider the Poisson multiset estimation since it is easier to relate with distribution estimation and is related to the Poissonization technique considered earlier.

6.1 Poisson Multiset Estimation

Let Λ be a list of k Poisson distributions, indexed $i = 1, 2, \ldots, k$, whose means are $\Lambda \stackrel{\text{def}}{=} (\Lambda(1), \Lambda(2), \ldots, \Lambda(k))$. Let $\overline{X} = X(1), X(2), \ldots, X(k)$ be samples drawn independently according to Λ, *i.e.,* $X(i) \sim \text{poi}(\Lambda(i))$, for $i = 1, 2, \ldots, k$. Let the collection of nonzero samples in $\overline{X}$ be denoted by

$$\mu(\overline{X}) \stackrel{\text{def}}{=} \overline{\mu} \stackrel{\text{def}}{=} \{\mu_1, \mu_2, \ldots, \mu_m\} \stackrel{\text{def}}{=} \{X(i) : X(i) > 0, 1 \leq i \leq k\},$$

where $m \stackrel{\text{def}}{=} m(\overline{X})$ is the number of nonzero samples and $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$. The information in $\mu(\overline{X})$ is equivalently conveyed using prevalences

$$\varphi_\mu \stackrel{\text{def}}{=} \varphi_\mu(\overline{X}) \stackrel{\text{def}}{=} \varphi_\mu(\overline{\mu}) \stackrel{\text{def}}{=} |\{i : \mu_i = \mu, 1 \leq i \leq m\}|,$$

the number of samples whose value is μ, for $\mu \in \{1, 2, \ldots\}$, and the profile of $\overline{X}$, given by

$$\varphi(\overline{X}) \stackrel{\text{def}}{=} \varphi(\overline{\mu}) \stackrel{\text{def}}{=} \overline{\varphi} \stackrel{\text{def}}{=} (\varphi_1, \varphi_2, \ldots).$$

In general, for every $\overline{\varphi}$, we associate a unique $\overline{\mu} = \mu(\overline{\varphi})$ in which the number of μ is φ_μ. Thus, there is a 1-1 correspondence between profiles $\overline{\varphi}$ and sample collections $\overline{\mu}$, and we use both with the same meaning.

Given a sample collection $\overline{\mu} \sim \Lambda$, we want to estimate the Poisson multiset

$$\mathcal{M}(\Lambda) \stackrel{\text{def}}{=} (\lambda_1, \lambda_2, \ldots, \lambda_k) \stackrel{\text{def}}{=} \{\Lambda(1), \Lambda(2), \ldots, \Lambda(k)\}$$

of the unknown distributions Λ. Without loss of generality, we assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. We often use Λ to imply $\mathcal{M}(\Lambda)$ for brevity, whenever it is clear from the context. In general, we are not given k or any other information about Λ.

A Poisson multiset estimator Q outputs a multiset of nonnegative reals $Q_{\overline{\mu}} \stackrel{\text{def}}{=} \{q_1, q_2, \ldots, q_{k'}\}$ as an estimate of $\mathcal{M}(\Lambda)$, given input $\overline{\mu} \sim \Lambda$. In particular, the empirical or sequence maximum likelihood estimator Q^{emp} outputs $Q^{\text{emp}}_{\overline{\mu}} = \overline{\mu}$ due to the following reasoning. The probability that Λ produces $\overline{x}$ is

$$\Lambda(\overline{x}) \stackrel{\text{def}}{=} \Pr\left(\overline{X} = \overline{x}\right) = \prod_{i=1}^{k} \text{poi}(\Lambda(i), x(i)),$$

where $\overline{X} \sim \Lambda$. Thus, the Λ that maximizes the probability of observing $\overline{x}$ such that $x(i) = \mu_i$ for $i = 1, \ldots, m$ and $x(i) = 0$ for $m < i \leq k$ is $\hat{\Lambda}_{\overline{X}} \stackrel{\text{def}}{=} \arg\max_{\Lambda} \Lambda(\overline{x}) = (\mu_1, \ldots, \mu_m, 0, \ldots, 0)$, and $\mathcal{M}(\hat{\Lambda}_{\overline{x}}) = \overline{\mu}$.

To measure the quality of estimators, one may consider any suitable distance measure between the multisets $\mathcal{M}(\Lambda)$ and $Q = Q_{\overline{\mu}}$. A natural choice for such a distance, motivated by the related problem of distribution estimation, is the *sorted L_1 distance*, simply referred to as the L_1 distance, between Λ and Q, and is given by

$$|\Lambda - Q| \stackrel{\text{def}}{=} |\mathcal{M}(\Lambda) - Q| \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} |\lambda_i - q_i|,$$

where the means of Λ and Q are arranged in decreasing order, *i.e.,* $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_k \geq 0$ and $\lambda_i = 0$ for $i > k$, and similarly for Q. The following is a simple example of the various definitions above.

Example 56. Let $\Lambda = (\Lambda(1), \Lambda(2), \Lambda(3), \Lambda(4)) = (3.5, 3, 1.1, 0.2)$. Let $X(1) = 3, X(2) = 4, X(3) = 1, X(4) = 0$ and thus, $\overline{\mu} = 4, 3, 1$. The empirical estimate is therefore $Q = Q_{\overline{\mu}} = \overline{\mu} = \{4, 3, 1\}$. And $|\Lambda - Q| = 0.5 + 0 + 0.1 + 0.2 = 0.8$. □

Unlike the empirical estimator that considers the likelihood of observing a specific sequence $\overline{x}$ whose profile is $\mu(\overline{x}) = \overline{\mu}$, it is natural to consider the overall likelihood of $\overline{\mu}$, *i.e.,* the probability of observing any sequence whose sample collection is $\overline{\mu}$. The likelihood of a sample collection $\overline{\mu}$, or equivalently its profile $\overline{\varphi}$, under Λ is

$$\begin{aligned}\Lambda(\overline{\varphi}) &\stackrel{\text{def}}{=} \Pr\left(\varphi(\overline{X}) = \overline{\varphi}\right) = \sum_{\overline{x}:\varphi(\overline{x})=\overline{\varphi}} \Lambda(\overline{x}) \\ &\stackrel{\text{def}}{=} \Pr\left(\mu(\overline{X}) = \overline{\mu}\right) = \sum_{\overline{x}:\mu(\overline{x})=\overline{\mu}} \Lambda(\overline{x}),\end{aligned}$$

where $\overline{X} \sim \Lambda$. A simple enumeration of all $\overline{x}$ that have the same profile $\overline{\varphi}$ leads to an explicit expression for $\Lambda(\overline{\varphi})$ as given in Observation 57. For all $\overline{\mu}$, let

$$S_{\overline{\mu}} \stackrel{\text{def}}{=} S_{\overline{\varphi}} \stackrel{\text{def}}{=} \sum_{i=1}^{m} \mu_i \stackrel{\text{def}}{=} \sum_{\mu=1}^{\infty} \varphi_\mu \cdot \mu$$

denote the sum of its samples, also referred to as the length of the corresponding profile $\overline{\varphi}$. Following the notation in Section 2.4, notice that $\overline{\varphi} \in \Phi^{S_{\overline{\varphi}}}$. Similarly, for all Λ, let

$$S_\Lambda \stackrel{\text{def}}{=} \sum_{i=1}^{k} \lambda_i$$

be the sum of its means.

Observation 57. *For all Λ such that $\mathcal{M}(\Lambda) = (\lambda_1, \ldots, \lambda_k)$ and for all $\overline{\mu} = (\mu_1, \ldots, \mu_m)$ such that $\varphi(\overline{\mu}) = \overline{\varphi} = (\varphi_1, \varphi_2, \ldots)$,*

$$\begin{aligned}\Lambda(\overline{\varphi}) = \Lambda(\overline{\mu}) &= \frac{1}{\prod_{\mu=1}^{\infty} \varphi_\mu!} \sum_{\sigma \in [k]^{\underline{m}}} \prod_{i=1}^{m} \text{poi}\big(\lambda_{\sigma(i)}, \mu(i)\big) \\ &= \text{poi}(S_\Lambda, S_{\overline{\varphi}}) \cdot \frac{S_{\overline{\varphi}}!}{\prod_{\mu=1}^{\infty} (\mu!)^{\varphi_\mu} \varphi_\mu!} \sum_{\sigma \in [k]^{\underline{m}}} \prod_{i=1}^{m} \left(\frac{\lambda_{\sigma(i)}}{S_\Lambda}\right)^{\mu(i)}. \qquad \square\end{aligned}$$

The maximum likelihood of a profile $\overline{\varphi}$ over all Λ is

$$\hat{\Lambda}(\overline{\varphi}) \stackrel{\text{def}}{=} \max_{\Lambda} \Lambda(\overline{\varphi}) = \hat{\Lambda}_{\overline{\varphi}}(\overline{\varphi}),$$

where

$$\hat{\Lambda}_{\overline{\varphi}} = \arg\max_{\Lambda} \Lambda(\overline{\varphi})$$

is the maximizing distribution. When restricted to a class of Poisson multisets $\mathcal{L}$, we define the corresponding maximum likelihoods and maximizing distributions

$$\hat{\Lambda}_{\mathcal{L}}(\overline{\varphi}) \stackrel{\text{def}}{=} \max_{\Lambda \in \mathcal{L}} \Lambda(\overline{\varphi}) \quad \text{and} \quad \hat{\Lambda}_{\mathcal{L},\overline{\varphi}} \stackrel{\text{def}}{=} \arg\max_{\Lambda \in \mathcal{L}} \Lambda(\overline{\varphi}).$$

The profile maximum likelihood (PML) estimator Q^{PML} simply outputs the Poisson multiset $Q^{\text{PML}}_{\overline{\mu}} = \hat{\Lambda}_{\overline{\varphi}}$ or $\hat{\Lambda}_{\mathcal{L},\overline{\varphi}}$ corresponding to input $\overline{\mu}$, *i.e.*, $\overline{\varphi} = \varphi(\overline{\mu})$.

Using the general result on the competitivity of ML for distribution estimation in Lemma 40, we show that the PML estimator is competitive for Poisson multiset estimation in the next lemma. The additional ingredient needed to bound $|\mathcal{Z}|$ in Lemma 40, other than the fact that $|\Phi^n| \le e^{3\sqrt{n}}$, is that $S_{\overline{\mu}}$ concentrates around S_Λ as given by the following observation.

Observation 58. *For all $0 < \epsilon < 1$ and all Λ such that $S_\Lambda \ge \frac{1}{\epsilon^2(1-\epsilon)}$, if $\overline{\mu} \sim \Lambda$, then*

$$\Pr\left(|S_{\overline{\mu}} - S_\Lambda| \ge \epsilon S_\Lambda\right) \le 2e^{-\epsilon^2 S_\Lambda/3}.$$

For all Λ such that $S_\Lambda \ge 2$,

$$\Pr\left(S_{\overline{\mu}} > 2S_\Lambda\right) \le e^{-S_\Lambda/6}.$$

Proof. Since $S_{\overline{\mu}} \sim \text{poi}(S_\Lambda)$, the result follows from Observation 7. □

Lemma 59. *Let $\mathcal{L}$ be a collection of Poisson multisets Λ such that $S_\Lambda \ge 2$. Let $D(\cdot,\cdot)$ be a distance measure on $\mathcal{L} \times \mathcal{L}$. Suppose there exists an estimator Q such that for some $\epsilon, \delta > 0$ and for all $\Lambda \in \mathcal{L}$, when $\overline{\mu} \sim \Lambda$,*

$$\Pr\left(D(\Lambda, Q_{\overline{\mu}}) \ge \epsilon\right) \le \delta.$$

Then, the PML estimator $\hat{\Lambda}_{\mathcal{L},\varphi(\overline{\mu})}$ has error

$$\Pr\left(D(\Lambda, \hat{\Lambda}_{\mathcal{L},\varphi(\overline{\mu})}) \ge 2\epsilon\right) \le \delta e^{4\sqrt{S_\Lambda}} + e^{-S_\Lambda/6}.$$

Proof. The proof is similar to that of Lemma 41 for distribution multiset estimation. If $\overline{\mu} \sim \Lambda \in \mathcal{L}$, then the error probability of the PML estimator is

$$\begin{aligned}
&\Pr\left(D(\Lambda,\hat{\Lambda}_{\mathcal{L},\varphi(\overline{\mu})}) \ge 2\epsilon\right)\\
&\qquad \le \Pr\left(S_{\overline{\mu}} > 2S_\Lambda\right) + \Pr\Big(\left(D(\Lambda, \hat{\Lambda}_{\mathcal{L},\varphi(\overline{\mu})}) \ge 2\epsilon\right) \wedge (S_{\overline{\mu}} \le 2S_\Lambda)\Big)\\
&\qquad \le\; e^{-S_\Lambda/6} + \delta e^{4\sqrt{S_\Lambda}}.
\end{aligned}$$

In the last inequality, the bound on the first term follows from Observation 58. For the second term, we notice that for the Q in the lemma's statement which has a small error $\Pr\Big(\left(D(\Lambda, Q_{\overline{\mu}}) \ge \epsilon\right) \wedge (S_{\overline{\mu}} \le 2S_\Lambda)\Big) \le \Pr\left(D(\Lambda, Q_{\overline{\mu}}) \ge \epsilon\right) \le \delta$. Hence, the result follows from the general result for competitivity of ML in Lemma 40, along with the fact that the number of $\overline{\mu}$ such that $S_{\overline{\mu}} \le 2S_\Lambda$ is at most

$2S_\Lambda \cdot |\Phi^{2S_\Lambda}| \leq e^{4\sqrt{S_\Lambda}}$ using Lemma 1. □

The above competitivity result is useful when $\delta < e^{-4\sqrt{S_\Lambda}}$. We show the existence of such multiset estimators Q that approximate Λ to within an L_1 distance of ϵS_Λ, where $\epsilon > 0$, with error probability $e^{-S_\Lambda^{0.9}}$, when the number of elements in Λ is $k = o(\epsilon^{2.1} n \log(n))$. We do this by showing that good distribution multiset estimators can be used to construct good Poisson multiset estimators, both under L_1 distance estimation guarantees. Hence, we use the distribution estimator in [68] to obtain a multiset estimator with the stated error guarantees. The simple idea behind this is that estimating a Poisson multiset Λ to within an L_1 distance of ϵS is almost equivalent to estimating the distribution multiset

$$\frac{\Lambda}{S_\Lambda} \stackrel{\text{def}}{=} \Big(\frac{\Lambda(1)}{S_\Lambda}, \frac{\Lambda(2)}{S_\Lambda}, \cdots, \frac{\Lambda(k)}{S_\Lambda}\Big)$$

to within a sorted L_1 distance of ϵ. The construction and equivalence are given by the following definition and lemma.

Definition 60. Let $\mathcal{L}$ be a class of Poisson multisets Λ and let $\mathcal{P} \stackrel{\text{def}}{=} \{\Lambda/S_\Lambda : \Lambda \in \mathcal{L}\}$ be the collection of the class of corresponding normalized distributions. Let $\widetilde{Q} \stackrel{\text{def}}{=} \widetilde{Q}_{\overline{\varphi}}$ be a profile-based distribution multiset estimator for $\mathcal{P}$. Then, the corresponding Poisson multiset estimator Q^{poi}, when given input $\overline{\mu}$ outputs

$$Q^{\text{poi}}_{\overline{\mu}} \;\stackrel{\text{def}}{=}\; S_{\overline{\mu}} \cdot \widetilde{Q}_{\varphi(\overline{\mu})},$$

where $\widetilde{Q}_{\varphi(\overline{\mu})}$ is the output of $\widetilde{Q}$ corresponding to input $\overline{\varphi} = \varphi(\overline{\mu}) \in \Phi^{S_{\overline{\mu}}}$. □

Lemma 61. *For $\epsilon \in (0,2)$, let $\mathcal{L}$ be a class of Poisson multisets Λ such that $S_\Lambda \geq \frac{8}{\epsilon^2(2-\epsilon)}$, and let $\mathcal{P} \stackrel{\text{def}}{=} \{\Lambda/S_\Lambda : \Lambda \in \mathcal{L}\}$. Let $\widetilde{Q}$ be a profile-based distribution estimator such that for all $P \in \mathcal{P}$, and all $n \geq \min\{S_\Lambda/2 : \Lambda \in \mathcal{L}\}$, when $\overline{Y} \sim P^n$,*

$$\Pr\big(|P - \widetilde{Q}_{\varphi(\overline{Y})}|_1 > \epsilon\big) \;\leq\; \delta(n),$$

where δ decreases monotonically in n. Then, the corresponding Poisson multiset estimator Q^{poi} (in Definition 60) is such that for all $\Lambda \in \mathcal{L}$, given $\overline{\mu} \sim \Lambda$,

$$\Pr\big(|\Lambda - Q^{\text{poi}}_{\overline{\mu}}| > 2\epsilon S_\Lambda\big) \;\leq\; \delta\big(S_\Lambda/2\big) + e^{-S_\Lambda/12} + 2e^{-\epsilon^2 S_\Lambda/3}.$$

Proof. Consider any $\Lambda \in \mathcal{L}$ and let $\overline{\mu} \sim \Lambda$. Then,

$$\begin{aligned}
&\Pr\big(|\frac{\Lambda}{S_\Lambda} - \widetilde{Q}_{\varphi(\overline{\mu})}| \le \epsilon\big) \\
&\qquad \le \Pr\Big(\big(|\frac{\Lambda}{S_\Lambda} - \widetilde{Q}_{\varphi(\overline{\mu})}| \le \epsilon\big) \wedge \big(S_{\overline{\mu}} \ge \frac{S_\Lambda}{2}\big)\Big) + \Pr\big(S_{\overline{\mu}} \le \frac{S_\Lambda}{2}\big) \\
&\qquad \le \delta\big(S_\Lambda/2\big) + e^{-S_\Lambda/12},
\end{aligned}$$

where the last inequality is due to the following reasoning. The discussion on Poissonization in Section 2.6 which implies that the distribution of $\varphi(\overline{\mu}) \in \Phi^{S_{\overline{\mu}}}$ is equivalent to that of $\varphi(\overline{Y})$ where $\overline{Y} \sim P^{n'}$, $P = \frac{\Lambda}{S_\Lambda}$ and $n' = S_{\overline{\mu}} \sim \text{poi}(S_\Lambda)$. Hence, if $n' = S_{\overline{\mu}} \ge \frac{S_\Lambda}{2}$, since δ is monotonically decreasing $\Pr(|\frac{\Lambda}{S_\Lambda} - \widetilde{Q}_{\varphi(\overline{\mu})}| \ge \epsilon) \le \delta(S_\Lambda/2)$, which explains the first term. The second term follows from the fact that $n' = S_{\overline{\mu}} \le \frac{S_\Lambda}{2}$ with probability $e^{-S_\Lambda/12}$ by Poisson tail bounds of Observation 7.

By Observation 58, we also have that

$$\Pr\big(|S_\Lambda - S_{\overline{\mu}}| \ge \epsilon S_\Lambda\big) \le 2e^{-\epsilon^2 S_\Lambda/3}.$$

To combining both these observations, we note that if $|\frac{\Lambda}{S_\Lambda} - \widetilde{Q}_{\varphi(\overline{\mu})}| \le \epsilon$ and $|S_\Lambda - S_{\overline{\mu}}| \le \epsilon S_\Lambda$, then (by a trivial abuse of notation)

$$\begin{aligned}
|\Lambda - Q^{\text{poi}}_{\overline{\mu}}| &= |S_\Lambda \cdot \frac{\Lambda}{S_\Lambda} - S_{\overline{\mu}} \cdot \widetilde{Q}_{\varphi(\overline{\mu})}| \\
&= |S_\Lambda \cdot \frac{\Lambda}{S_\Lambda} - S_\Lambda \cdot \widetilde{Q}_{\varphi(\overline{\mu})} + S_\Lambda \cdot \widetilde{Q}_{\varphi(\overline{\mu})} - S_{\overline{\mu}} \cdot \widetilde{Q}_{\varphi(\overline{\mu})}| \\
&\le S_\Lambda |\frac{\Lambda}{S_\Lambda} - \widetilde{Q}_{\varphi(\overline{\mu})}| + |S_\Lambda - S_{\overline{\mu}}| \\
&\le 2\epsilon S_\Lambda.
\end{aligned}$$

Hence, by union bound,

$$\begin{aligned}
\Pr\big(|\Lambda - Q^{\text{poi}}_{\overline{\mu}}| > 2\epsilon S_\Lambda\big) &\le \Pr\big(|\frac{\Lambda}{S_\Lambda} - \widetilde{Q}_{\varphi(\overline{\mu})}| \le \epsilon\big) + \Pr\big(|S_\Lambda - S_{\overline{\mu}}| \ge \epsilon S_\Lambda\big) \\
&\le \delta\big(S_\Lambda/2\big) + e^{-S_\Lambda/12} + 2e^{-\epsilon^2 S_\Lambda/3}.
\end{aligned}$$

□

We thus have the following corollaries.

Corollary 62. *Let $Q^{\text{poi,VV}}$ be the Poisson multiset estimator Q^{poi} corresponding to the distribution estimator $\widetilde{Q} = Q^{\text{VV}}$. For all $\epsilon > 0$, and for all Λ with sufficiently* *large S_Λ and $k = \mathcal{O}(\epsilon^{2.1} S_\Lambda \log(S_\Lambda))$, if $\overline{\mu} \sim \Lambda$, then*

$$\Pr\left(|\Lambda - Q_{\overline{\mu}}^{\text{poi,VV}}| \ge \epsilon S_\Lambda\right) \le e^{-S_\Lambda^{0.85}}$$

and

$$\Pr\left(|\Lambda - \hat{\Lambda}_{\varphi(\overline{\mu})}| \ge 2\epsilon S_\Lambda\right) \le e^{-S_\Lambda^{0.8}},$$

where the PML is restricted over multisets of support $\mathcal{O}(\epsilon^{2.1} S_\Lambda \log(S_\Lambda))$.

Proof. Follows from Lemma 61, Corollary 50 and Lemma 59. □

Corollary 63. *For all $\epsilon \in (0, \frac{1}{2})$, and for all Λ with sufficiently large S_Λ and $k = \mathcal{O}(\epsilon^{2.1} S_\Lambda)$, if $\overline{\mu} \sim \Lambda$, then*

$$\Pr\left(|\Lambda - Q_{\overline{\mu}}^{\text{emp}}| \ge \epsilon S_\Lambda\right) \le e^{-\epsilon^2 S_\Lambda/17}$$

and

$$\Pr\left(|\Lambda - \hat{\Lambda}_{\varphi(\overline{\mu})}| \ge 2\epsilon S_\Lambda\right) \le e^{-\epsilon^2 S_\Lambda/18},$$

where the PML is restricted over multisets of support $\mathcal{O}(\epsilon^{2.1} S_\Lambda)$.

Proof. Follows from Lemma 61, Lemma 38 and Lemma 59. □

We conclude this section on Poisson multiset estimation with the following remarks. Similar to the distribution estimation problem, Lemma 59 can be stated as competitive sample complexity result using the following observation and lemma.

Observation 64. *Let $\mathcal{L}$ be a collection of Λ that correspond to Poisson processes Λ' observed for time T,* i.e., *$\Lambda = T \cdot \Lambda'$. Let $\mathcal{L}' = \{\Lambda/T : \Lambda \in \mathcal{L}\}$. Suppose there is an estimator Q' for Λ' such that for some distance metric $D(\cdot,\cdot)$ on Λ', and some $\epsilon > 0$, when $\overline{\mu} \sim T \cdot \Lambda'$, $\Pr\left(D(\Lambda', Q'_{\overline{\mu}}) \ge \epsilon\right) \le \delta < \frac{1}{4}$. Then there exists an estimator Q'' for Λ' such that for all positive integers r, when $\overline{\mu} \sim (2r+1)T \cdot \Lambda'$, $\Pr\left(D(\Lambda', Q''_{\overline{\mu}}) \ge 3\epsilon\right) \le (4\delta)^r$.*

Proof Sketch. Similar to Observation 37. □

Lemma 65. *Following the notation in Observation 64, let $\mathcal{L}'$ be a collection of Λ' and suppose there is an estimator Q' for Λ' such that when $\overline{\mu} \sim T \cdot \Lambda'$, $\Pr\big(D(\Lambda', Q'_{\overline{\mu}}) \geq \epsilon\big) \leq \delta < \frac{1}{4}$. Then the PML estimator takes as input $\overline{\mu} \sim T' \cdot \Lambda'$ and outputs $Q^{\text{PML}} = \frac{1}{T'}\hat{\Lambda}_{\overline{\mu}}$ and has error $\Pr\big(D(\Lambda', Q^{\text{PML}}) \geq 6\epsilon\big) \leq \delta$, where $T' = \mathcal{O}\big(\max\{T, \frac{S_{\Lambda'}T^2}{\log^2(\frac{1}{4\delta})}\}\big)$.*

Proof Sketch. Using Observation 64, $r = \mathcal{O}\big(\max\{1, \frac{S_{\Lambda'}T}{\log^2(\frac{1}{4\delta})}\}\big)$ suffices to guarantee the existence of an estimator whose error probability is $\delta' = (4\delta)^r \leq \delta^2 e^{-10\sqrt{S'_{\Lambda}T'}}$ using $\overline{\mu} \sim T'\Lambda'$ where $T' = (2r+1)T$. Hence, by Lemma 59, the error probability of PML estimator is at most δ when $\overline{\mu} \sim T'\Lambda'$. □

Note that the sample complexity guarantees hold even when $\delta \geq e^{-4\sqrt{S_{\Lambda}}}$, unlike in Lemma 59. Lastly, we note the following simple relationship between the PML estimators for distribution multiset and Poisson multiset estimation.

Observation 66. *For all $\overline{\mu}$,*

$$\hat{\Lambda}_{\varphi(\overline{\mu})} = S_{\overline{\mu}} \cdot \hat{P}_{\varphi(\overline{\mu})}.$$

□

Thus, the computation of PML for Poisson multiset estimation is the same, *i.e.,* requires solving the same optimization problem, as that for the corresponding distribution multiset estimation. And all the earlier machinery used for computing PML for distribution multiset estimation is therefore directly applicable for Poisson multiset estimation as well.

6.2 Bernoulli Multiset Estimation

Let B be a list of k Bernoulli 0-1 distributions, indexed $i = 1, 2, \ldots, k$, whose success probabilities are $B \overset{\text{def}}{=} (\theta(1), \theta(2), \ldots, \theta(k))$. For a positive integer n and for each of $i = 1, 2, \ldots, k$, let $\overline{X}(i) \overset{\text{def}}{=} X(i,1), X(i,2), \ldots, X(i,n)$ be a sequence of n samples drawn independently according to bernoulli$(\theta(i))$, *i.e.,* $X(i,j)$ takes values 1 and 0 with probabilities $\theta(i)$ and $\overline{\theta}(i) \overset{\text{def}}{=} 1 - \theta(i)$ respectively, for $j = 1, \ldots, n$.

Let

$$\overline{\overline{X}} \overset{\text{def}}{=} \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_m \overset{\text{def}}{=} \left\{ \overline{X}(i) : \sum_{j=1}^{n} X(i,j) > 0, 1 \leq i \leq k \right\}$$

be the set of sequences $\overline{X}(i)$ in which at least one of the $X(i,j) = 1$ for $j = 1, 2, \ldots, n$. Here $m \overset{\text{def}}{=} m(\overline{\overline{X}})$ is the number of $\overline{X}(i)$ that have at least one 1.

Given a sample collection $\overline{\overline{X}} \sim B$, we want to find the Bernoulli multiset

$$\mathcal{M}(B) \overset{\text{def}}{=} (\theta_1, \theta_2, \ldots, \theta_k) \overset{\text{def}}{=} \{\theta(1), \theta(2), \ldots, \theta(k)\}$$

of the unknown distributions B. Without loss of generality, we assume $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_k$. We often use B to imply $\mathcal{M}(B)$ for brevity, whenever it is clear from the context. In general, we are not given k or any other information about B. An estimator Q outputs a multiset of probabilities $Q_{\overline{\overline{X}}} \overset{\text{def}}{=} \{q_1, q_2, \ldots, q_{k'}\}$ corresponding to each possible input $\overline{\overline{X}}$.

A simple example of Q is the empirical estimator $Q^{\text{emp}}_{\overline{\overline{X}}} \overset{\text{def}}{=} \{\frac{\mu_1}{n}, \ldots, \frac{\mu_m}{n}\}$ where μ_i is the number of 1's in $\overline{X}_i$, for $i = 1, 2, \ldots, m$. Notice that for a given B and $\overline{\overline{X}}$, the probability of observing $\overline{X}(1), \ldots, \overline{X}(k)$ such that $\overline{X}(i) = \overline{X}_i$ for $i = 1, \ldots, m$ and $\overline{X}(i) = 0, 0, \ldots, 0$ for $i = m+1, \ldots, k$ is

$$B\big(\overline{X}(i), \ldots, \overline{X}(k)\big) = \prod_{i=1}^{k} \theta(i)^{\mu_i} (1 - \theta(i))^{n - \mu_i},$$

where $\mu_i = 0$ for $i = m+1, \ldots, k$. Hence, $Q^{\text{emp}}_{\overline{\overline{X}}} = \hat{B}_{\overline{X}(1), \ldots, \overline{X}(k)}$ maximizes the likelihood of observing $\overline{X}(1), \ldots, \overline{X}(k)$.

To evaluate the performance of estimators, one may consider various distance measures between the underlying multiset B and estimated multiset $Q = Q_{\overline{\overline{X}}}$. A natural choice for such a distance, motivated from the related distribution estimation problem, is the *sorted* L_1 *distance*, or simply the L_1 distance, between B and Q and is defined as

$$|B - Q| \overset{\text{def}}{=} |\mathcal{M}(B) - Q| \overset{\text{def}}{=} \sum_{i=1}^{\infty} |\theta_i - q_i|,$$

where the probabilities of B and Q are arranged in decreasing order. The following example illustrates the problem.

Example 67. Let $B = (\theta(1), \ldots, \theta(5)) = (\frac{1}{8}, 1, \frac{1}{2}, \frac{1}{8}, \frac{1}{8})$. Then its multiset is $\mathcal{M}(B) = (1, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$. Suppose $n = 3$ samples are taken from each of these distributions and the sample sequences obtained are $\overline{X}(1) = (0, 0, 0)$, $\overline{X}(2) = (1, 1, 1)$, $\overline{X}(3) = (1, 0, 1)$, $\overline{X}(4) = (0, 0, 0)$, $\overline{X}(5) = (1, 0, 0)$. We are given only the sequences $\overline{\overline{X}} = \overline{X}(2)$, $\overline{X}(3)$, $\overline{X}(5)$ that contain at least one 1 to estimate B. The empirical estimator outputs $Q = Q^{\text{emp}}_{\overline{\overline{X}}} = (1, \frac{2}{3}, \frac{1}{3})$. If in addition, we are given that $k = 5$ and that each of the $\theta(i)$ have a uniform prior over $[0, 1]$, then one obtains the *Laplace* or *add-one* estimate for each of the processes as $(\frac{3+1}{3+2}, \frac{2+1}{3+2}, \frac{1+1}{3+2}, \frac{0+1}{3+2}, \frac{0+1}{3+2})$ and outputs $Q' = (\frac{4}{5}, \frac{3}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5})$. The multisets $Q'' = (1, \frac{1}{2})$, $Q''' = (1, 1, \frac{1}{3})$. are also allowed estimates, although it is clear that they cannot generate the given collections. We still notice that $|P - Q''| = \frac{3}{8}$ is smaller than $|P - Q| = \frac{5}{8}$. □

Since we want to estimate $\mathcal{M}(B)$ and the sequences are generated *i.i.d.*, it is natural to consider the profile of $\overline{\overline{X}}$, which conveys the multiset of counts of 1 in the different sequences in $\overline{\overline{X}}$. The profiles considered here are similar to those considered for distribution multiset and Poisson multiset estimation. For $i = 1, 2, \ldots, m$, let $\overline{X}_i \overset{\text{def}}{=} X_{i,1}, \ldots, X_{i,n}$ and μ_i, the multiplicity of $\overline{X}_i$, is the number of appearances of 1 in $\overline{X}_i$, *i.e.,*

$$\mu_i \overset{\text{def}}{=} \mu(\overline{X}_i) \overset{\text{def}}{=} \sum_{j=1}^{n} X_{i,j}.$$

For $\mu = 1, \ldots, n$, the prevalence of μ is

$$\varphi_\mu \overset{\text{def}}{=} |\{i : \mu_i = \mu, 1 \le i \le m\}| \overset{\text{def}}{=} \sum_{i=1}^{m} \mathbb{1}_{[\mu_i = \mu]},$$

the number of sequences $\overline{X}_i$ that have μ 1's. The profile of $\overline{\overline{X}}$ is

$$\varphi(\overline{\overline{X}}) \overset{\text{def}}{=} \overline{\varphi} \overset{\text{def}}{=} (\varphi_1, \varphi_2, \ldots, \varphi_n).$$

The profile $\varphi(\overline{\overline{X}})$ is equivalently conveyed as the multiplicity vector

$$\mu(\overline{\overline{X}}) \overset{\text{def}}{=} \mu(\overline{\varphi}) \overset{\text{def}}{=} \overline{\mu} \overset{\text{def}}{=} \{\mu_1, \mu_2, \ldots, \mu_m\}.$$

Without loss of generality, we assume that $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_m$ and that $\overline{X}_i$'s are indexed in decreasing order of their multiplicities. For every $\overline{\mu}$, there is a unique

$\overline{\varphi} = \varphi(\overline{\mu})$ and vice versa. We use both $\overline{\varphi}$ and $\overline{\mu}$ with the same meaning. For all $\overline{\varphi}$ and its corresponding $\overline{\mu}$, let

$$S_{\overline{\varphi}} \overset{\text{def}}{=} S_{\overline{\mu}} \overset{\text{def}}{=} \sum_{i=1}^{m} \mu_i \overset{\text{def}}{=} \sum_{\mu=1}^{\infty} \varphi_\mu \cdot \mu.$$

Following the notation in Section 2.4, it is implied that $\overline{\varphi} \in \Phi^{S_{\overline{\varphi}}}$.

The likelihood of a profile $\overline{\varphi}$ under Bernoulli distributions B is the probability of observing a sample collection $\overline{\overline{X}}$ whose profile is $\overline{\varphi}$ under B , given by

$$B(\overline{\varphi}) \ \overset{\text{def}}{=} B\big(\varphi(\overline{\overline{X}}) = \overline{\varphi}\big) \ \overset{\text{def}}{=} \sum_{\overline{x}(1),\ldots,\overline{x}(k):\varphi(\overline{\overline{x}})=\overline{\varphi}} B(\overline{x}(1), \ldots, \overline{x}(k)).$$

Since the probability of $\overline{x}(1), \ldots, \overline{x}(k)$ under B is

$$B(\overline{x}(1), \ldots, \overline{x}(k)) = \prod_{i=1}^{k} (\theta(i))^{\mu(\overline{x}_i)} (1 - \theta(i))^{n-\mu(\overline{x}_i)},$$

a simple enumeration of all $\overline{x}(1), \ldots, \overline{x}(k)$ such that the corresponding $\overline{\overline{x}}$ has profile $\varphi(\overline{\overline{x}}) = \overline{\varphi}$ leads to the following explicit expression for $B(\overline{\varphi})$.

Observation 68. *For all B such that $\mathcal{M}(B) = (\theta_1, \ldots, \theta_k)$ and for all $\overline{\varphi} = (\varphi_1, \varphi_2, \ldots)$ such that $\mu(\overline{\varphi}) = (\mu_1, \ldots, \mu_m)$,*

$$\begin{aligned} B(\overline{\varphi}) &= \frac{1}{(k-m)! \prod_{\mu=1}^{n} \varphi_\mu!} \sum_{\sigma \in S_k} \prod_{i=1}^{k} \binom{n}{\mu_i} \theta_{\sigma(i)}^{\mu_i} (1 - \theta_{\sigma(i)})^{n-\mu_i} \\ &= \frac{(n!)^m}{\prod_{\mu=1}^{n} (\mu!(n-\mu)!)^{\varphi_\mu} \varphi_\mu!} \sum_{\sigma \in [k]^{\underline{m}}} \prod_{i=1}^{m} \theta_{\sigma(i)}^{\mu_i} (1 - \theta_{\sigma(i)})^{n-\mu_i}. \qquad \square \end{aligned}$$

The maximum likelihood of a profile $\overline{\varphi}$ over all B is

$$\hat{B}(\overline{\varphi}) \overset{\text{def}}{=} \max_{B} \Lambda(\overline{\varphi}) = \hat{B}_{\overline{\varphi}}(\overline{\varphi}),$$

where

$$\hat{B}_{\overline{\varphi}} = \arg\max_{B} B(\overline{\varphi})$$

is the maximizing distribution. When restricted to a class of Bernoulli multisets $\mathcal{B}$, we define the corresponding maximum likelihoods and maximizing distributions

$$\hat{B}_{\mathcal{B}}(\overline{\varphi}) \overset{\text{def}}{=} \max_{B \in \mathcal{B}} B(\overline{\varphi}) \ \text{ and } \ \hat{B}_{\mathcal{B},\overline{\varphi}} \overset{\text{def}}{=} \arg\max_{B \in \mathcal{B}} B(\overline{\varphi}).$$

The profile maximum likelihood (PML) estimator Q^{PML} outputs the Bernoulli multiset $Q^{\text{PML}}_{\overline{\overline{X}}} = \hat{B}_{\varphi(\overline{\overline{X}})}$ or $\hat{B}_{\mathcal{B},\varphi(\overline{\overline{X}})}$ corresponding to input $\overline{\overline{X}}$.

Using the general result on the competitivity of ML for distribution estimation in Lemma 40, we show that the PML estimator is competitive for Bernoulli multiset estimation. Without loss of generality, since we want to estimate $\mathcal{M}(B)$ and the sequences in $\overline{\overline{X}}$ are generated *i.i.d.*, we only consider estimators Q that depend on $\overline{\overline{X}}$ only through its profile $\varphi(\overline{\overline{X}})$, *i.e.*, $Q_{\overline{\overline{X}}} = Q_{\varphi(\overline{\overline{X}})}$, by an argument similar to subsection 3.2.1 or [9, Section 3.1.3]. To apply the general competitivity result, we needed to bound $|\mathcal{Z}|$ in Lemma 40. For this, we show that for all B and large n, when $\overline{\overline{X}} \sim B$, $S_{\varphi(\overline{\overline{X}})}$ concentrates around nS_B where

$$S_B \overset{\text{def}}{=} \sum_{i=1}^{k} \theta_i$$

is the sum of its success probabilities.

Observation 69. *For all $\epsilon \in (0,1)$ and all B, if $\overline{\overline{X}} \sim B$, then*

$$\Pr\left(|S_{\varphi(\overline{\overline{X}})} - nS_B| \geq \epsilon n S_B\right) \leq 2e^{-\epsilon^2 S_B/3}$$

and

$$\Pr\left(S_{\overline{\mu}} > 2nS_B\right) \leq e^{-S_B/3}.$$

Proof. Using Chernoff bounds from Fact 8 on the quantity $S_{\varphi(\overline{\overline{X}})} = \sum_{i=1}^{m} \mu_i = \sum_{i=1}^{k}\sum_{j=1}^{n} X(i,j)$, which is a sum of 0-1 independent random variables and has mean $E[\sum_i \sum_j X(i,j)] = \sum_i n\theta(i) = nS_B$. □

Lemma 70. *Let $\mathcal{B}$ be a collection of Bernoulli multisets B and $D(\cdot,\cdot)$ be a distance measure on $\mathcal{B} \times \mathcal{B}$. Suppose there exists a profile-based estimator Q such that for some $\epsilon, \delta > 0$ and for all $B \in \mathcal{B}$, when $\overline{\overline{X}} \sim B$,*

$$\Pr\left(D(\Lambda, Q_{\varphi(\overline{\overline{X}})}) \geq \epsilon\right) \leq \delta.$$

Then, the PML estimator $\hat{B}_{\mathcal{B},\varphi(\overline{\overline{X}})}$ has error

$$\Pr\left(D(\Lambda, \hat{B}_{\mathcal{B},\varphi(\overline{\mu})}) \geq 2\epsilon\right) \leq \delta e^{4\sqrt{S_B}} + e^{-S_B/3}.$$

Proof. Similar to the proof of Lemma 41 for distribution multiset estimation and that of Lemma 59 for Poisson multiset estimation, if $\overline{\overline{X}} \sim B \in \mathcal{B}$, then the error probability of the PML estimator is

$$\begin{aligned}
&\Pr\left(D(B,\hat{B}_{\mathcal{B},\varphi(\overline{\overline{X}})}) \geq 2\epsilon\right) \\
&\qquad \leq \Pr\left(S_{\varphi(\overline{\overline{X}})} > 2nS_B\right) + \Pr\Big(\left(D(\Lambda, \hat{\Lambda}_{\mathcal{L},\varphi(\overline{\overline{X}})}) \geq 2\epsilon\right) \wedge \left(S_{\varphi(\overline{\overline{X}})} \leq 2nS_B\right)\Big) \\
&\qquad \leq \; e^{-nS_B/3} + \delta e^{4\sqrt{nS_B}}.
\end{aligned}$$

In the last inequality, the bound on the first term follows from Observation 69. For the second term, the Q in the statement of the lemma has a small error given by $\Pr\Big(\left(D(\Lambda, Q_{\varphi(\overline{\overline{X}})}) \geq \epsilon\right) \wedge \left(S_{\varphi(\overline{\overline{X}})} \leq 2S_B\right)\Big) \leq \Pr\left(D(\Lambda, Q_{\varphi(\overline{\overline{X}})}) \geq \epsilon\right) \leq \delta$. Hence, the result follows from the general result for competitivity of ML in Lemma 40, along with the fact that the number of $\overline{\varphi}$ such that $S_{\overline{\varphi}} \leq 2nS_B$ is at most $2nS_B \cdot |\Phi^{2nS_B}| \leq e^{4\sqrt{nS_B}}$ using Lemma 1. □

To make use of the above lemma, we show the existence of multiset estimators Q whose error probability is at most $\exp(-4\sqrt{nS_B})$ for non-trivial choices of $\mathcal{B}$ and $D(\cdot,\cdot)$. We do this by relating the Bernoulli multiset estimation problem to that of distribution estimation problem when L_1 distance is used for evaluating the quality of estimation. In the process, we show that good distribution multiset estimators can be easily used as good Bernoulli multiset estimators. The simple idea behind this is that estimating a Bernoulli multiset B to within an L_1 distance of $\epsilon n S_B$ is essentially equivalent to estimating the distribution multiset

$$\frac{B}{S_B} \stackrel{\text{def}}{=} \left(\frac{\theta(1)}{S_B}, \frac{\theta(2)}{S_B}, \ldots, \frac{\theta(k)}{S_B}\right)$$

to within a sorted L_1 distance of ϵ. The complete construction is stated below.

Definition 71. Let $\mathcal{B}$ be a class of Bernoulli multisets B and let $\mathcal{P} \stackrel{\text{def}}{=} \{B/S_B : B \in \mathcal{B}\}$ be the collection of the class of corresponding normalized distributions. Let $\widetilde{Q} \stackrel{\text{def}}{=} \widetilde{Q}_{\overline{\varphi}}$ be a profile-based distribution multiset estimator for $\mathcal{P}$. Then, the corresponding Bernoulli multiset estimator Q^{bern} for input $\overline{\overline{X}} \sim B \in \mathcal{B}$ is defined as follows. For $i = 1, \ldots, m$, generate independent $n_i \sim \text{poi}(\frac{n}{2})$. If some $n_i > n$, terminate the estimation process and output error. Otherwise, for each of $i = 1, \ldots, m$, let $\overline{Y}_i$ consist of first n_i samples of $\overline{X}_i$, *i.e.,* $\overline{Y}_i = X_{i,1}, X_{i,2}, \ldots, X_{i,n_i}$.

Let $\mu_i' \stackrel{\text{def}}{=} \mu(\overline{Y}_i)$ be the number of 1's in $\overline{Y}_i$ for $i = 1, 2, \ldots, m$. And let $\overline{\varphi}' = (\varphi_1', \varphi_2', \ldots)$ be the profile of $\overline{\mu}' \stackrel{\text{def}}{=} \{\mu_i' : \mu_i' > 0, 1 \le i \le m\}$. In other words, $\varphi_\mu' = \sum_{i=1}^m \mathbb{1}_{[\mu_i' = \mu]}$ is the number of μ in $\overline{\mu}'$. Then, output

$$Q^{\text{bern}}_{\varphi(\overline{\overline{\overline{X}}})} \stackrel{\text{def}}{=} \frac{S_{\varphi(\overline{\overline{\overline{X}}})}}{n} \cdot \widetilde{Q}_{\overline{\varphi}'}. \qquad \square$$

The following lemma shows that if $\widetilde{Q}$ is a good distribution estimator, then the corresponding Q^{bern} is good Bernoulli multiset estimator, both in L_1 distance.

Lemma 72. *For $\epsilon \in (0, 2)$, let $\mathcal{B}$ be a class of distribution multisets B such that $S_\Lambda \ge \frac{8}{\epsilon^2(2-\epsilon)}$, and let $\mathcal{P} \stackrel{\text{def}}{=} \{B/S_B : B \in \mathcal{B}\}$. Let $\widetilde{Q}$ be a profile-based distribution estimator such that for all $P \in \mathcal{P}$, and all $\ell \ge n \cdot \min\{S_B/4 : B \in \mathcal{B}\}$, when $\overline{Y} \sim P^\ell$,*

$$\Pr\left(|P - \widetilde{Q}_{\varphi(\overline{Y})}|_1 > \epsilon\right) \le \delta(\ell),$$

where δ decreases monotonically in ℓ. Then, the corresponding Bernoulli multiset estimator Q^{bern} (in Definition 71) is such that for all $B \in \mathcal{B}$, given $\overline{\overline{\overline{X}}} \sim B$,

$$\Pr\left(|B - Q^{\text{bern}}_{\varphi(\overline{\overline{\overline{X}}})}| > 2\epsilon S_B\right) \le \delta\big(nS_B/4\big) + 2e^{-\epsilon^2 n S_B/3} + e^{-nS_B/24} + ke^{-n/12}.$$

Proof. Consider any $B \in \mathcal{B}$ and let $\overline{\overline{\overline{X}}} \sim B$. And consider the intermediate steps of Definition 71 involved in obtaining Q^{bern} from the $\widetilde{Q}$ in the lemma statement. By Poisson tail bounds of Observation 7, and union bound, probability that some $n_i > n$ for $i = 1, \ldots, m$, in Definition 71 is at most $me^{-n/12} \le ke^{-n/12}$.

If all $n_i < n$, by the discussion on Poissonization in Section 2.6, all $\mu_i' \sim \text{poi}(n\theta_i/2) = \text{poi}\big((nS_B/2) \cdot (\theta_i/S_B)\big)$. (Notice that while we did not explicitly take care of symbols that did not appear, hypothetically generating $n_i \sim \text{poi}(n/2)$ for $i = m+1, \ldots, k$ and taking first $n/2$ samples would have still resulted in $\mu_i' = 0$ for $i = m+1, \ldots, k$, as is the case currently.) It follows that $\overline{\varphi}'$ is the profile of $\overline{Z} \sim (\frac{B}{S_B})^{\text{poi}(nS_B/2)}$. Hence $\overline{\varphi}'$ has length $S_{\overline{\varphi}'} = \sum_\mu \varphi_\mu' \mu \ge nS_B/4$ with probability $\ge 1 - e^{-nS_B/24}$ by Poisson tail bounds. In that case, $\Pr\left(|\frac{B}{S_B} - \widetilde{Q}_{\overline{\varphi}'}|_1 \ge \epsilon\right) \le \delta(S_{\overline{\varphi}'}) \le \delta(nS_B/4)$.

Using Observation 69, $\Pr\big(|S_{\varphi(\overline{\overline{X}})} - nS_B| \ge \epsilon n S_B\big) \le 2e^{-\epsilon^2 n S_B/3}$. Similar to the proof of Lemma 61, if $|\frac{B}{S_B} - \widetilde{Q}_{\overline{\varphi}'}|_1 \le \epsilon$ and $|S_{\varphi(\overline{\overline{X}})} - nS_B| \le \epsilon n S_B$, then

$$|B - Q^{\text{bern}}_{\varphi(\overline{\overline{X}})}| = |S_B \cdot \frac{B}{S_B} - \frac{S_{\varphi(\overline{\overline{X}})}}{n}\widetilde{Q}_{\overline{\varphi}'}| \le 2\epsilon S_B.$$

Assembling the facts above, and using union bound for bounding the overall error probability,

$$\begin{aligned}
\Pr\big(|B - \widetilde{Q}_{\varphi(\overline{\overline{X}})}| > 2\epsilon S_B\big) &\le \Pr\big(n_i > n \text{ for some } i = 1, \dots, m\big) \\
&\quad + \Pr\big(S_{\overline{\varphi}'} < nS_B/4\big) \\
&\quad + \Pr\Big(\big(|\frac{B}{S_B} - \widetilde{Q}_{\overline{\varphi}'}|_1 \ge \epsilon\big) \wedge \big(S_{\overline{\varphi}'} \ge nS_B/4\big)\Big) \\
&\quad + \Pr\big(|S_{\varphi(\overline{\overline{X}})} - nS_B| \ge \epsilon n S_B\big) \\
&\le ke^{-n/12} + e^{-nS_B/24} + \delta\big(nS_B/4\big) + 2e^{-\epsilon^2 n S_B/3}.
\end{aligned}$$ □

The following corollaries are applications of the competitivity result for the PML-based estimator and the method for converting distribution multiset estimators to Bernoulli multiset estimators.

Corollary 73. *Let $Q^{\text{bern,VV}}$ be the Bernoulli multiset estimator Q^{bern} corresponding to the distribution estimator $\widetilde{Q} = Q^{\text{VV}}$. For all $\epsilon > 0$, for all sufficiently large n, and for all B with sufficiently large nS_B and $k = \mathcal{O}(\epsilon^{2.1} nS_B \log(nS_B))$, if $\overline{\overline{X}} \sim B$, then*

$$\Pr\big(|B - Q^{\text{bern,VV}}_{\varphi(\overline{\overline{X}})}| \ge \epsilon S_B\big) \le e^{-(nS_B)^{0.85}}$$

and

$$\Pr\big(|B - \hat{B}_{\varphi(\overline{\overline{X}})}| \ge 2\epsilon S_B\big) \le e^{-(nS_B)^{0.8}}.$$

where the PML is restricted over multisets of support $\mathcal{O}(\epsilon^{2.1} nS_B \log(nS_B))$.

Proof. Follows from Lemma 72, Corollary 50 and Lemma 70. □

Corollary 74. *For all $\epsilon \in (0, \frac{1}{2})$, for sufficiently large n, and for all B with sufficiently large nS_B and $k = \mathcal{O}(\epsilon^{2.1} nS_B)$, if $\overline{\overline{X}} \sim B$, then*

$$\Pr\big(|B - Q^{\text{emp}}_{\overline{\overline{X}}}| \ge \epsilon S_B\big) \le e^{-\epsilon^2 n S_B/17}$$

and

$$\Pr\left(|B - \hat{B}_{\varphi(\overline{\overline{X}})}| \geq 2\epsilon S_B\right) \leq e^{-\epsilon^2 n S_B/18},$$

where the PML is restricted over multisets of support $\mathcal{O}(\epsilon^{2.1} n S_B)$.

Proof. Follows from Lemma 72, Lemma 38 and Lemma 70. □

Lemma 70 can be stated as competitive sample complexity result using the following observation and lemma. The sample complexity guarantee holds even when $\delta \geq e^{-4\sqrt{nS_B}}$, unlike in Lemma 70.

Observation 75. *Let $\mathcal{B}$ be a collection of B. Suppose there is an estimator Q for $\mathcal{B}$ such that for some distance metric $D(\cdot,\cdot)$ on $\mathcal{B}$, and some $\epsilon > 0$, when $\overline{\overline{X}} \sim B \in \mathcal{B}$ and sequences in $\overline{\overline{X}}$ are of length n, $\Pr\left(D(B, Q_{\overline{\overline{X}}}) \geq \epsilon\right) \leq \delta < \frac{1}{4}$. Then there exists an estimator Q' for $\mathcal{B}$ such that for all positive integers r, when $\overline{\overline{X}}' \sim B$ and sequences in $\overline{\overline{X}}'$ are of length $n' = (2r+1)n$, $\Pr\left(D(B, Q'_{\overline{\overline{X}}'}) \geq 3\epsilon\right) \leq (4\delta)^r$.*

Proof Sketch. Similar to Observation 37. □

Lemma 76. *Let $\mathcal{B}$ be a collection of B and suppose there is a profile-based estimator Q for $\mathcal{B}$ such that when $\overline{\overline{X}} \sim B \in \mathcal{B}$, and sequences in $\overline{\overline{X}}$ are of length n, $\Pr\left(D(B, Q_{\varphi(\overline{\overline{X}})}) \geq \epsilon\right) \leq \delta < \frac{1}{4}$. Then the PML estimator takes as input $\overline{\overline{X}}' \sim B$, where sequences in $\overline{\overline{X}}'$ are of length n', and has error $\Pr\left(D(B, \hat{B}_{\varphi(\overline{\overline{X}}')}) \geq 6\epsilon\right) \leq \delta$, where $n' = \mathcal{O}\big(\max\{n, \frac{n^2 S_B}{\log^2(\frac{1}{4\delta})}\}\big)$.*

Proof Sketch. Using Observation 64, $r = \mathcal{O}\big(\max\{1, \frac{nS_B}{\log^2(\frac{1}{4\delta})}\}\big)$ suffices to guarantee the existence of an estimator whose error probability is $\delta' = (4\delta)^r \leq \delta^2 e^{-10\sqrt{n' S_B}}$ using $\overline{\overline{X}}' \sim B$ where sequences in $\overline{\overline{X}}'$ have length $n' = (2r+1)n$. Hence, by Lemma 70, the error probability of PML estimator is at most δ using $\overline{\overline{X}}'$. □

To summarize and conclude, we showed that PML estimator for Bernoulli multiset estimator is competitive with other estimators. We showed applications of this result by showing that good distribution multiset estimators can be used to construct good Bernoulli multiset estimators, which is a useful result in itself. However, we note from Observation 68 that the exact computation of PML for

Bernoulli multiset estimation involves a related but different optimization problem than that for distribution and Poisson multiset estimation. Several results analogous to that for PML computation for distribution estimation are shown in [1]. An EM algorithm and its experimental results for directly approximating PML for Bernoulli multiset estimation are shown in [2, 1]. It is similar to the EM algorithm for approximating PML for distribution estimation considered in [50, 46, 75].

Acknowledgement

Chapter 6 is adapted from Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, “Estimating multisets of Bernoulli processes”, Submitted to *IEEE Symposium on Information Theory (ISIT)*, 2012; and Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, “Estimating multiple processes”, In preparation, 2012; The dissertation author is a primary researcher and author of these papers.

Bibliography

[1] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Estimating multiple processes. *In preparation*, 2012.

[2] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Estimating multisets of bernoulli processes. In *Submitted to IEEE Symposium on Information Theory (ISIT)*, 2012.

[3] J. Acharya, H. Das, A. Orlitsky, and S. Pan. Algebraic computation of pattern maximum likelihood. In *Proceedings of IEEE Symposium on Information Theory (ISIT)*, pages 400–404, 2011.

[4] 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 (ISIT)*, pages 1493–1497, 2010.

[5] J. Acharya, A. Orlitsky, and S. Pan. The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns. In *Proceedings of IEEE Symposium on Information Theory (ISIT)*, pages 1135 –1139, 2009.

[6] J. Acharya, A. Orlitsky, and S. Pan. Recent results on pattern maximum likelihood. In *IEEE Information Theory Workshop on Networking and Information Theory (ITW)*, pages 251–255, 2009.

[7] 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.

[8] 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.

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

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

[11] J. Bunge and M. Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88:364–373, 1993.

[12] Ana Cardoso-Cachopo. Datasets for single-label text categorization.

[13] A. Chao. Nonparametric estimation of the number of classes in a population. *Scandanavian Journal of Statistics: Theory and Applications*, 11:265–270, 1984.

[14] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *ACL*, pages 310–318, 1996.

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

[16] David A. Cox, John Little, and Donal O'Shea. *Using Algebraic Geometry*. Springer, 2004.

[17] David A. Cox, John Little, and Donal O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag New York, 2007.

[18] I. Csiszar and J. Korner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akademiai Kiado, 1979.

[19] Imre Csiszár. The method of types. *IEEE Transactions on Information Theory*, 44(6):2505–2523, 1998.

[20] L.D. Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19(6):783–795, November 1973.

[21] A.K. Dhulipala and A. Orlitsky. Universal compression of markov and related sources over arbitrary alphabets. *IEEE Transactions on Information Theory*, 53:4182–4190, 2006.

[22] David L. Donoho and Richard C. Liu. Geometrizing rates of convergence. *Annals of Statistics*, 19:633–701, 1991.

[23] M. Feder and N. Merhav. Universal composite hypothesis testing: A competitive minimax approach. *IEEE Transactions on Information Theory*, 48:1504–1517, 2002.

[24] R. Fisher, A. Corbet, and C. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12:42–48, 1943.

[25] W.A. Gale, K.W. Church, and D. Yarowsky. A method for disambiguating word senses. *Computers and Humanities*, 26:415–419, 1993.

[26] I.J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, December 1953.

[27] I.J. Good. Turing's anticipation of Empirical Bayes in connection with the cryptanalysis of the Naval Enigma. *Journal of Statistics Computation and Simulation*, 66:101–111, 2000.

[28] I.J. Good and G.H. Toulmin. The number of new species and the increase in population coverage when the sample is increased. *Biometrika*, 43(1):45–63, 1956.

[29] Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Sublinear estimation of entropy and information distances. *ACM Transactions on Algorithms*, 5(4), 2009.

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

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

[32] Bruno M. Jedynak and Sanjeev Khudanpur. Maximum likelihood set for estimating a probability mass function. *Neural Computation*, 17:1–23, 2005.

[33] S.M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400–401, March 1987.

[34] R. Keener, E. Rothman, and N. Starr. Distributions on partitions. *Annals of Statistics*, 15(4):1466–1481, 1987.

[35] 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 (ISIT)*, pages 1478–1482, 2010.

[36] P.S. Laplace. *Philosphical essays on probabilities.* Springer Verlag, New York, Translated by A. Dale from the 5th (1825) edition, 1995.

[37] Lucien LeCam. *Asymptotic Methods in Statistical Decision Theory.* Springer, New York, 1986.

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

[39] David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pages 4–15, 1998.

[40] Dinesh Manocha and John F. Canny. Multipolynomial resultant algorithms. *Journal of Symbolic Computation*, 15:99–122, 1993.

[41] C. Mao and B. Lindsay. A poisson model for the coverage problem with a genomic application. *Biometrika*, 89:669–682, 2002.

[42] D. McAllester and R. Schapire. On the convergence rate of Good Turing estimators. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

[43] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996.

[44] 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.

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

[46] A. Orlitsky, Sajama, N.P. Santhanam, K. Viswanathan, and J. Zhang. Algorithms for modeling distributions over large alphabets. In *Proceedings of IEEE Symposium on Information Theory (ISIT)*, pages 304–304, 2004.

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

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

[49] 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.

[50] Alon Orlitsky, Narayana P. Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *UAI '04*, pages 426–435, 2004.

[51] Liam Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004.

[52] Liam Paninski. Variational minimax estimation of discrete distributions under kl loss. In *NIPS*, 2004.

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

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

[55] 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.

[56] Y. Ritov and P. J. Bickel. Achieving information bounds in non- and semiparametric models. *Annals of Statistics*, 18:925–938, 1990.

[57] H.E. Robbins. Estimating the total probability of the unobserved outcomes of an experiment. *Annals of Mathematical Statistics*, 39:256–257, 1968.

[58] 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.

[59] J. Shtarkov. Coding of discrete sources with unknown statistics. In I. Csiszár and P. Elias, editors, *Topics in Information Theory (Coll. Math. Soc. J. Bolyai, no. 16)*, pages 559–574. Amsterdam, The Netherlands: North Holland, 1977.

[60] H.S. Sichel. The GIGP distribution model with applications to physics literature. *Czechoslovak Journal of Physics, Ser. B*, 36:133–137, 1986.

[61] Fei Song and W. Bruce Croft. A general language model for information retrieval. In *CIKM*, pages 316–321, 1999.

[62] W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission*, 34(2):142–146, 1998.

[63] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences.* Wiley, 2001.

[64] W. Szpankowski and M. Weinberger. Minimax redundancy for large alphabets. In *Proceedings of IEEE Symposium on Information Theory (ISIT)*, pages 1488–1492, 2010.

[65] W. Szpankowski and M. Weinberger. Minimax redundancy for large alphabets, 2010.

[66] R. Thisted and B. Efron. Estimating the number of unseen species: How many words did shakespeare know. *Biometrika*, 63:435–447, 1976.

[67] Gregory Valiant and Paul Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:180, 2010.

[68] Gregory Valiant and Paul Valiant. Estimating the unseen: An $n/\log(n)$-sample estimator for entropy, support size, and other distribution properties, with a proof of optimality via two new central limit theorems. In *STOC '11: Proceedings of the 42nd annual ACM symposium on Theory of computing*, 2011.

[69] Gregory Valiant and Paul Valiant. The power of linear estimators. In *FOCS '11: Proceedings of the 52nd Annual Symposium on Foundations of Computer Science*, pages 403–412, 2011.

[70] 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.

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

[72] Aaron B. Wagner, Pramod Viswanath, and Sanjeev R. Kulkarni. Probability estimation in the rare-events regime. *IEEE Transactions on Information Theory*, 57(6):3207–3229, 2011.

[73] Q. Xie and A.R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.

[74] D. Zelterman. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference*, 18:225–237, 1988.

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

[76] G. K. Zipf. *Human behavior and the principle of least effort: An Introduction to Human Ecology*. Addison-Wesley, Cambridge, USA, 1949.

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