On estimating the probability multiset

A. Orlitsky∗ N.P. Santhanam† K. Viswanathan† J. Zhang§

June 12, 2011

PRELIMINARY DRAFT

Abstract

Many statistical properties are determined not by the probabilities of the possible outcomes, but just by the multiset of these probabilities. We estimate this probability multiset from an observed data sample. Akin to standard maximum likelihood, which maximizes the probability of the observed outcomes, we propose profile maximum likelihood (PML), which maximizes the probability of the observed profile—the number of symbols appearing any given number of times. We prove the existence and consistency of the PML estimate, establish some of its modeling and functional properties, and derive it in closed form for several simple and short profiles.

1 Introduction

1.1 Overview

Distributions associate a probability with each possible outcome. For example, a coin may turn heads with probability 2/3 and tails with probability 1/3. Yet many statistical properties are insensitive to the values of the outcomes and are determined by just the multiset of probabilities irrespective of the outcomes they are associated with. For example, the probability that the above coin will turn the same value in two independent flips is determined by just the probability set \{\frac{2}{3}, \frac{1}{3}\} regardless of the association between the faces and these two probabilities.

Value-insensitive properties include many distribution characteristics such as support size, entropy, and diversity indices. They also encompass sample properties such as the expected number of distinct, new, or unlikely, outcomes that will appear in a sample of a given size, or the size of a new sample required to observe a certain number, or fraction, of the possible outcomes.

These properties therefore arise in a variety of scientific and engineering applications. Biologists may seek the number of species in an environment, immunologists might assess the number virus strains responsible for a certain percentage of infections, pharmacologists could estimate the number of new compounds expected to be found in a drug-discovery experiment.

∗ECE and CSE, UC San Diego, alon@ucsd.edu
†EECS, UC Berkeley, prasad@talk.ucsd.edu
‡HP Labs, krishnamurthy.viswanathan@hp.com
§Duke University, junan.zhang@duke.edu
of a given size, and engineers might evaluate the probability that a computer cache will not contain requested memory pages, or that demand for a resource will exceed its supply.

All the quantities above are determined by just the multiset of probabilities, regardless of the outcomes they are associated with. For example, the probability multiset \{.3,.2,.2,.1,.1,.1\} corresponds to six possible outcomes, three of which cover 70% of the probability, and \(6 - .7^{10} - 2 \cdot .8^{10} - 3 \cdot .9^{10} \approx 4.71\) of which are expected to be encountered in 10 independent samples. We are therefore interested in estimating this probability multiset based on a sample of observed data.

A tempting two-step approach for estimating the probability multiset is to first find the maximum-likelihood estimate of the element probabilities, and then ignore the association between the elements and their probabilities. We call the resulting multiset the standard, or sequence, maximum likelihood (SML) estimate.

Maximum-likelihood is among the most popular distribution-estimation techniques. Assuming a class of potential distributions, it postulates that a sample is generated by the distribution maximizing its probability. For example, upon observing a sequence of 30 heads and 70 tails generated by independent coin flips, maximum likelihood would propose that \(p(\text{heads}) \overset{\text{def}}{=} p = 0.3\) and \(p(\text{tails}) \overset{\text{def}}{=} 1 - p = .7\), the values maximizing the sequence probability \(p^{30}(1 - p)^{70}\). This natural estimate also coincides with the intuitive empirical-frequency estimate that assigns to each outcome a probability corresponding to the fraction of times it appeared. Ignoring the association between the faces and their probabilities, the SML estimate for the coin-flip sequence is therefore the probability set \{.7,.3\}.

SML performs well when each possible outcome appears enough times to reliably estimate its probability. This typically occurs when the sample size is large relative to the support size. For instance when as in the coin example above 100 samples were used to estimate just two probabilities.

Yet in other cases, SML may fall short. Consider sampling the DNA sequences of \(n\) individuals. Clearly, all \(n\) sequences will be distinct, and SML would hypothesize a uniform distribution over the set of \(n\) observed sequences, resulting in the SML estimate \(\{\frac{1}{n}, \ldots, \frac{1}{n}\}\) of the probability multiset. However, this estimate does not capture the essential property that no two sequences are identical. A distribution over a much larger, essentially infinite, range of outcomes, will better explain the data.

We propose an alternative estimation method. It is based on the simple premise that since we seek only the multiset of probabilities, not the probability of specific elements, the estimate should not depend on the specific values seen, but only on the observed pattern, reflecting the co-occurrences of elements. Furthermore, since we estimate the multiset from an observed sample, assumed as usual to consist of iid drawings, the order of the elements does not matter.

The sample’s pattern can therefore in turn be replaced by its profile, the number of elements appearing any given number of times. For example, in the coin-flip sequence, one element was observed 30 times, and the other, 70.

It is therefore the probability of the sample’s pattern, or equivalently profile, that we maximize, and we call the method pattern-, or profile- maximum likelihood (PML). In this and two subsequent papers we introduce PML and show that it coincides with SML when the sample size is large relative to the number of possible outcomes, but better models the data for smaller sample sizes.

This paper motivates and defines the PML estimate and establishes its existence, consistency,
and modeling properties. The second paper determines the PML estimate for some simple and short data samples, and establishes some of its properties for more complex samples. The third paper describes algorithms for computing the PML estimate and presents experimental results indicating its efficacy even when the sample size is significantly smaller than the support set.

1.2 Examples

We introduce pattern and profile maximum likelihood via three examples. The first and simplest consists of just three samples. It is perhaps surprising that even with this smallest nontrivial sample size some of PML’s benefits can be observed.

Suppose that the first three independent and identically distributed (i.i.d.) samples from an unknown distribution are \(@ @ \). Clearly \( p(\@) = 2/3 \) and \( p(\#) = 1/3 \) maximize the sequence probability at \( 4/27 \). Hence the SML estimate of the probability multiset is \( \{ 2, 3 \} \).

However, the appearance of the specific symbols @ and #, while increasing their future reappearance probability, should not affect the estimate of the probability multiset. This estimate should remain unchanged if any two other symbols appear instead, hence must depend only on the sequence’s pattern where the first two elements are identical and the third is different. It is the probability of this pattern, not of the observed values, that should be maximized.

Under the \( \{ 2, 3 \} \) multiset suggested by sequence maximum-likelihood, the probability of this, first two same, third different, pattern is

\[
\left( \frac{2}{3} \right)^2 \cdot \frac{1}{3} + \frac{2}{3} \cdot \left( \frac{1}{3} \right)^2 = \frac{2}{9}.
\]

However, a \( \{ \frac{1}{2}, \frac{1}{2} \} \) distribution is more likely to generate this pattern as it assigns it probability

\[
\left( \frac{1}{2} \right)^2 \cdot \frac{1}{2} + \frac{1}{2} \cdot \left( \frac{1}{2} \right)^2 = \frac{1}{4}.
\]

It is easy to see that \( \{ \frac{1}{2}, \frac{1}{2} \} \) maximizes the probability of the first two same, third different, pattern over all discrete distributions, not just with two, but with any number of support elements. Any such distribution \( \{ p_1, p_2, \ldots \} \) assigns to this pattern probability of

\[
\sum_i p_i^2 (1 - p_i) = \sum_i p_i \cdot [p_i (1 - p_i)] \leq \sum_i p_i \cdot \frac{1}{4} = \frac{1}{4}.
\]

We will shortly show that \( \{ \frac{1}{2}, \frac{1}{2} \} \) maximizes the pattern’s probability also for non-discrete distributions. Hence \( \{ \frac{1}{2}, \frac{1}{2} \} \) is the PML estimate of the first two same, third different, pattern.

Comparing the \( \{ \frac{2}{3}, \frac{1}{3} \} \) SML and the \( \{ \frac{1}{2}, \frac{1}{2} \} \) PML multisets, we observe that PML better explains the data even for such a small sample size. Suppose for simplicity, as both estimates suggest, that the underlying distribution consists of two elements. Three samples cannot result in two elements appearing the same number (1.5) of times. Hence the appearance of one element twice and the other once is the closest possible to a uniform distribution that can be observed from three samples. Furthermore, with two support elements and three samples, either all observations equal each other, or, as was observed, two equal each other and one differs. Any distribution differing from \( \{ \frac{1}{2}, \frac{1}{2} \} \) will increase the probability of three equal observations, hence decrease the probability of the observed two-and-one outcome.
The mathematical difference between the SML and PML multiset estimates of @@@# is simple. SML maximizes the probability of the observed sequence of values. It therefore postulates a distribution \((p, q)\), where \(p\) is the probability of @ and \(q\) is the probability of #, and maximizes \(p^2q\). PML by contrast maximizes the probability of the observed pattern. For binary distributions for example it therefore specifies only the multiset \([p, q]\), leaving open the question as to which of \(p\) and \(q\) is associated with @ and #, and maximizes \(p^2q + q^2p\).

For @@@#, the SML and PML estimates \(\{\frac{1}{2}, \frac{1}{2}\}\) and \(\{\frac{2}{3}, \frac{1}{3}\}\), consist of different probabilities, but agree that the distribution is over two elements. For some data samples, the estimated support sizes differs too. Imagine that the first 20 samples from an unknown distribution consist of 10 distinct symbols, each appearing twice. Maximum likelihood would maximize the probability of the observed values, namely the product

\[
\prod_a P^2(a)
\]

taken over the 10 observed symbols, and would therefore hypothesize the uniform distribution over them. On the other hand, a uniform distribution over \(k \geq 10\) symbols assigns to the observed pattern, and to every other pattern whose profile consists of 10 symbols, each appearing twice, a probability of

\[
\frac{k^{10}}{k^{20}},
\]

where \(k^\underline{m} = k!/\,(k-m)!\) is the \(m\)-th falling power of \(k\). This probability is maximized by \(k = 12\), and results in the second part of this paper show that it exceeds the probability assigned to the pattern by any other distribution.

Therefore, while SML suggests that the distribution underlying this 20-element sequence has as many symbols as those observed, PML proposes a distribution over a larger set. This support increase better agrees with the intuition that in a sample whose size is comparable to the number of support symbols, some fraction of the symbols will not appear.

The discrepancy between PML and SML grows as the number of observed symbols increases relative to the sample size. In the limit, consider again the \(n\) DNA sequences. SML assigned probability \(1/n\) to each previously observed sequence and 0 to all others. By contrast, PML would observe that DNA sequences have the simple pattern where each of \(n\) distinct symbols appears just once. To maximize this pattern’s probability, PML would suggest that the sequences are generated by a distribution over a larger, ideally infinite, alphabet. This estimate clearly much better agrees with our intuition.

For one implication of the difference between SML and PML, consider predicting properties of unseen data. Suppose that having observed the \(n\) DNA sequences, we try to estimate how many distinct sequences will be observed if we sample, say, \(10n\) new individuals, and how many of these sequences will differ from those seen in the first \(n\) samples. SML resulted in a uniform distribution over the \(n\) observed sequences. It would therefore suggest that the number of distinct sequences among the \(10n\) new samples will be at most \(n\), and that all the sequences will be among those observed in the original sample. By contrast, PML postulates a distribution over an infinite support, hence would suggest that all new sequences will be distinct, and all will differ from the original ones. Clearly, the latter will better match the new observations.

Finally, observe that since samples are typically assumed to be drawn \(i.i.d.\), the order in which the elements appear is irrelevant. The pattern probability is therefore determined by the data’s profile, the number of elements appearing any given number of times. For example, in the
first example, the sequences @@#, @#@@, and #@@ share the same profile where one element appears once and one element appears twice, hence any underlying distribution will assign the same probability to their respective patterns. The PML distribution maximizing the pattern probability is therefore also determined by just the data’s profile.

1.3 Experiments

To see how PML approximates the original distribution, consider three experiments. In the first, shown in Figure 1, a uniform distribution over 500 elements, represented by the solid blue, rectangular, line was used to generate 250 samples with replacement. In a typical sample, of the 500 elements, 6 appeared thrice, 45 appeared twice, 142 appeared once, and 307 did not appear at all.

SML would state that the distribution consists of 5+46+143=194 elements, 5 with probability 3/250, 46 with probability 2/250, and 143 with probability 1/250. This distribution is represented by the red, parallel-strokes, jagged line. SML therefore misses over 60% of the distribution’s elements, and its uniformity.

By contrast, the PML distribution is shown by the dashed green line. Although only 196 symbols were observed, appearing with unequal frequencies, and 306 elements did not, the PML distribution estimates the original distribution as essentially uniform over roughly 500 elements. We should note that the algorithm used to derive the PML estimate, described in the third paper in this series, is heuristic. Hence there is hope that the actual PML distribution estimates the original even better.

In an extension of this experiment, shown in Figure 2, 1000 independent samples were selected with replacement from the same uniform distribution over 500 elements. Now each element appears on the average four times, hence maximum likelihood, shown in the red parallel-strokes line, estimates the number of elements well. Yet it still misses the distribution’s uniformity. This time, the PML distribution, shown again by the dashed purple line, replicates the true distribution essentially exactly.

Since the PML approach does not assume a uniform, or any distribution, it can be applied to estimate any other distributions. Figure 3 shows a similar experiment with a Zipf distribution over 450 elements where for \( i = 51, \ldots, 500 \), \( p_i = c/i \), and \( c \) is a normalization factor. Sampling the distribution 500 times yields the SML estimate which again differs significantly from the true distribution and estimates a support of only about 280 elements. But PML approximates the true distribution quite well, and even though about 170 elements do not appear in the sample at all, it posits a sharp drop after about 450 elements.

One of the most important value-insensitive properties of a distribution is its support-size,
and a related well-studied problem is that of using an observed sample to estimate the number of distinct elements that will appear in a new sample of a given size.

Perhaps the most popular estimator for this task was devised by Good and Toulmin and is further described in the next section. Among others, it was used by Efron and Thisted to estimate the number of previously unobserved species, or more generally, symbols, that will appear in a new sample. They assumed that the number of butterflies of each species was distributed Poisson independently of all other species, and that the Poisson means were drawn independently according to a Gamma prior with a parameter chosen to best fit the data.

The Poisson model was further pursued by Efron and Thisted who lower bounded the expected number of previously unobserved species, or more generally, symbols, that will appear in a new sample. They expressed this lower bound in terms of the expected number of symbols that appear any given number of times in the observed sample. Since the expectations are unknown, they approximated them by the actual numbers observed. In 1987, a new poem presumed to have been written by Shakespeare was discovered. It contained 9 words not previously encountered in the Bard’s work—fairly close to the 7 predicted by these estimates.

Additional work on the Poisson model was undertaken in future samples, and greatly outperforms it for large ones.

2 Previous and related results

While we are not aware of research estimating specifically the probability multiset, two related directions for estimating distributions over large support have been studied extensively.

Value-insensitive distribution properties, especially the support size and the number of new elements expected to be observed in further samples have been considered for the better part of the last century.

In 1943, Fisher et al. used a modestly-sized sample to estimate the number of Malayan butterfly species and the expected number of species that will appear any given number of times if a new sample of a prescribed size was collected. They assumed that the number of butterflies of each species was distributed Poisson independently of all other species, and that the Poisson means were drawn independently according to a Gamma prior with a parameter chosen to best fit the data.

The Poisson model was further pursued by Efron and Thisted who lower bounded the expected number of previously unobserved species, or more generally, symbols, that will appear in a new sample. They expressed this lower bound in terms of the expected number of symbols that appear any given number of times in the observed sample. Since the expectations are unknown, they approximated them by the actual numbers observed. In 1987, a new poem presumed to have been written by Shakespeare was discovered. It contained 9 words not previously encountered in the Bard’s work—fairly close to the 7 predicted by these estimates. Additional work on the Poisson model was undertaken in future samples, and greatly outperforms it for large ones.
The number of times each low-probability element appears may be distributed Poisson. Yet the joint distribution of the number of times all elements appear is not independent Poisson, but rather, multinomial. Several subsequent studies in the Bayesian framework took this multinomial approach. They assume a prior on the alphabet size \( k \), and for each \( k \), assume a symmetric \( k \)-dimensional Dirichlet prior with parameter \( \alpha \), where \( \alpha \) is either assumed or estimated [?], or distributed according to another prior [?, ?]. These works are related to the one considered here as, given the assumed prior, the posterior distribution is a function of the observed profile.

The multinomial approach was used to estimate the underlying alphabet size [?, ?], the probability of unseen data [?], and most recently, to determine the size of a sample needed to ensure, with the given confidence level, that a certain fraction of the possible symbols will appear [?]. Additional approaches to the problem can be found in [?, ?, ?, ?, ?], and a review of these and other techniques appears in [?].

The more traditional problem where the correspondence between the support elements and their probabilities is of interest has of course been studied even longer. In the the 18th century, realizing the aforementioned shortcomings of maximum likelihood, Laplace [?] assigned to each observed and unobserved symbol a probability proportional to one more than the number of times it This Laplace, or add-one, and other add-constant estimators have been applied and studied extensively and shown to work well when the sample size is large compared to the underlying alphabet [?, ?, ?, ?].

However, when the sample size is small relative to the alphabet size, add-constant estimators do not work well [?]. Good and Turing encountered this problem when deciphering the Enigma code, also in the early 1940’s. British intelligence was in possession of the Kengruppenbuch, the German cipher book that contained all possible secret keys. Based on previously decrypted messages, Good and Turing tried to estimate the distribution of pages from which U-boat commanders chose their secret keys.

Good and Turing devised an eponymous estimator that evaluates the probabilities of each previously observed outcome and the combined probability of all yet unseen outcomes. The estimator is non-Bayesian and, though unintuitive, appeared to perform well in practice. The Good-Turing estimator was published by Good after the war [?, ?], and has since been incorporated into a variety of applications, such as speech recognition, e.g., [?, ?], spelling correction [?], and information retrieval [?]. On the theoretical side, several intuitive motivations for the estimator’s efficacy were given [?, ?, ?, ?], and some of its convergence properties were established [?].

It is easy to see that estimating the probability of each observed outcome and of all unseen outcomes combined is equivalent to estimating the probability of the sequence’s pattern. This pattern approach was used [?] to derive a natural variant of the Good-Turing estimator that asymptotically does not underestimates the probability of any sequence, and to obtain probability estimates whose probability estimates converge to the true one uniformly over all alphabets. This approach also showed [?] that while the compression redundancy of \( i.i.d. \) sequences over arbitrary alphabets is infinite, the corresponding redundancy of their patterns diminishes to zero.

In contrast to these works, we do not assume a prior, nor a particular parametrization of the underlying distribution. ??? add? ???
3 Preliminaries

Starting with sequences, we define patterns, profiles, and their probabilities.

3.1 Patterns and profiles

Let $\mathfrak{x} \equiv x^n \equiv x_1 \ldots x_n$ be a length-$n$ sequence. Then

\[ A(\mathfrak{x}) \equiv \{x_1, \ldots, x_n\}, \]

and

\[ m \equiv |A(\mathfrak{x})| \]

are the set and number of distinct symbols appearing in $\mathfrak{x}$. The index of $x \in A(\mathfrak{x})$ is

\[ i(x) \equiv \min \{|A(x_i)| : x_i = x\}, \]

one more than the number of distinct symbols preceding $x$’s first appearance in $\mathfrak{x}$. The pattern of $\mathfrak{x}$ is the concatenation

\[ \psi(\mathfrak{x}) \equiv i(x_1)i(x_2)\ldots i(x_n), \]

of all indices. For example, the sequence $\mathfrak{x} = “abracadabra”$ consists of $m = 5$ symbols, $a, b, r, c, d$. Their indices are $i(a) = 1$, $i(b) = 2$, $i(r) = 3$, $i(c) = 4$, and $i(d) = 5$, hence

\[ \psi(abracadabra) = 12314151231. \]

It is easy to see that a string of positive integers is a pattern iff the first occurrence of any $i \in \mathbb{P} \equiv \{1, 2, 3, \ldots\}$ precedes that of $i + 1$. For example, 1, 11, 12, and 1213 are patterns, while 2, 21, and 1312 are not.

If a symbol $\psi$ in a pattern repeats $i$ consecutive times, we abbreviate it by $\psi^i$. For example, representing the pattern 11222113 as $1^22^23^1$.

A pattern of the form $1^{\mu_1}2^{\mu_2}\ldots m^{\mu_m}$ where $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_m$, e.g., $1^22^33^41^1$, is canonical.

As mentioned earlier, the probability of a pattern is determined by its profile, defined next. The multiplicity of $x$ in $\mathfrak{x}$ is

\[ \mu_x \equiv \mu_x(\mathfrak{x}) \equiv |\{i : x_i = x\}|, \]

the number of times $x$ appears in $\mathfrak{x}$. The minimum multiplicity $\mu_{\min} \equiv \mu_{\min}(\mathfrak{x})$ of $\mathfrak{x}$ is the smallest multiplicity of any of its elements, and the maximum multiplicity $\mu_{\max} \equiv \mu_{\max}(\mathfrak{x})$ of $\mathfrak{x}$ is the largest multiplicity of any of its elements. The prevalence of an integer $\mu \in \mathbb{P}$ in $\mathfrak{x}$ is

\[ \varphi_\mu \equiv \varphi_\mu(\mathfrak{x}) \equiv |\{x : \mu_x = \mu\}|, \]

the number of symbols with multiplicity $\mu$, namely appearing $\mu$ times, in $\mathfrak{x}$. The profile of $\mathfrak{x}$ is the formal product

\[ \varphi(\mathfrak{x}) \equiv \varphi(\mathfrak{x}) \equiv \prod_{\mu=n}^{1} \mu^{\varphi_\mu} \equiv n^{\varphi_n}(n-1)^{\varphi_{n-1}}\ldots 2^{\varphi_2}1^{\varphi_1}, \]

of all multiplicities and their prevalences, where $\mu^0$ terms are typically omitted. For example, in the sequence abracadabra, the multiplicities are $\mu_a = 5$, $\mu_b = \mu_r = 2$, and $\mu_c = \mu_d = 1$, hence the minimum and maximum multiplicities are $\mu_{\min} = 1$ and $\mu_{\max} = 5$. The prevalences are $\varphi_1 = 2,$
$\varphi_2 = 2$, $\varphi_5 = 1$, and $\varphi_i = 0$ for all other $i$, and the profile is $\bar{\varphi} = (11)^0(10)^0 \ldots 6^05^14^03^02^21^2 = 5^12^21^2$.

Patterns, as sequences, have profiles too. It is easy to see that the profile of a sequence is the same as that of its pattern. For example, the pattern of \textit{abracadabra} is 12314151231 whose profile is also $5^12^21^2$.

Given a profile $\bar{\varphi}$, we let

$$\tilde{\Psi}_{\bar{\varphi}} \overset{\text{def}}{=} \{ \tilde{\psi} : \bar{\varphi}(\tilde{\psi}) = \bar{\varphi} \}$$

be the set of patterns of profile $\bar{\varphi}$. For every $\bar{\varphi}$, exactly one of the patterns in $\tilde{\Psi}_{\bar{\varphi}}$ is canonical, and is called the \textit{canonical pattern} of $\bar{\varphi}$. For example, the canonical pattern of the profile 512212 is 152341512.

### 3.2 Probabilities

We now define the distribution underlying the data and the probability it induces over sequences, patterns, and profiles. Since the distribution maximizing the probability of a pattern may have an arbitrarily large support, it is mathematically convenient to model parts of a distribution as continuous.

A \textit{discrete distribution} is a mapping $P : K \rightarrow [0, 1]$ where $K$ is a discrete set and $\sum_{a \in K} P(a) = 1$. A \textit{continuous distribution} is a mapping $P : \mathcal{X} \rightarrow [0, \infty)$ where $\mathcal{X}$ is an interval in the real line and $\int_{x \in \mathcal{X}} P(x)dx = 1$. A \textit{mixed distribution} is a weighted combination of a discrete distribution over a discrete set $K$ and a continuous distribution over an interval $\mathcal{X}$ such that $\sum_{a \in K} P(a) + \int_{\mathcal{X}} P(x)dx = 1$.

Suppose that $n \in \mathbb{P}$ data samples are collected sequentially from some distribution $P$. The observed samples will form an \textit{i.i.d.} sequence $\overline{X} = X^n = X_1X_2 \ldots X_n$ where the probability of $\overline{X} = x^n$ is

$$P(\overline{X}) \overset{\text{def}}{=} P(\overline{X} = \overline{x}) = \prod_{i=1}^n P(x_i).$$

The random sequence’s pattern

$$\tilde{\Psi} = \psi(\overline{X})$$

will be distributed according to

$$P(\tilde{\psi}) \overset{\text{def}}{=} P(\tilde{\Psi} = \tilde{\psi}) = P(\psi(\overline{X}) = \tilde{\psi}) = P(\{ \overline{x} : \psi(\overline{x}) = \tilde{\psi} \}),$$

the probability that a length-n sequence generated according to $P$ has pattern $\tilde{\psi}$.

For example, if $P$ is a discrete distribution over $\{a_1, a_2\}$ with $P(a_1) = p_1$ and $P(a_2) = p_2$, where $p_1 + p_2 = 1$, then for $n = 2$, the possible patterns are 11 and 12 and they occur with probability

$$P(\tilde{\Psi} = 11) = P(\{a_1a_1, a_2a_2\}) = p_1^2 + p_2^2 \quad \text{and} \quad P(\tilde{\Psi} = 12) = P(\{a_1a_2, a_2a_1\}) = 2p_1p_2.$$

By contrast, if $P$ is a continuous distribution over an interval $\mathcal{X}$ then, with probability one, two elements drawn according to $P$ will differ, hence

$$P(\tilde{\Psi} = 11) = P(\{xx : x \in \mathcal{X}\}) = 0 \quad \text{and} \quad P(\tilde{\Psi} = 12) = P(\{x_1x_2 : x_1 \neq x_2 \in \mathcal{X}\}) = 1.$$  

Finally, if $P$ is a mixed distribution over the union of $\{a_1, a_2\}$ and an interval $\mathcal{X}$, with $P(a_1) = p_1$, $P(a_2) = p_2$, and $P(\mathcal{X}) = q$, where $p_1 + p_2 + q = 1$, then

$$P(\tilde{\Psi} = 11) = P(\{a_1a_1, a_2a_2\}) + P(\{xx : x \in \mathcal{X}\}) = p_1^2 + p_2^2.$$
and

\[
P(\Psi = 12) = P(\{a_1a_2, a_2a_1\}) + P(\{a_1x, xa_1 : x \in X\}) + P(\{a_2x, xa_2 : x \in X\})
+ P(\{x_1x_2 : x_1 \neq x_2 \in X\})
= 2p_1p_2 + 2p_1q + 2p_2q + q^2
= 1 - p_1^2 - p_2^2.
\]

Note that the pattern probability is determined by the multiset \(\{P(a) : a \in K\}\) of discrete probabilities and the interval probability \(P(X)\). The values of the elements in \(K\) and \(X\), the arrangement of the probabilities over \(K\), and the precise distribution over \(X\) do not matter. Furthermore, \(P(X) = 1 - \sum_{a \in K} P(a)\), and is thus determined by the discrete probabilities. We therefore sometimes identify any distribution, discrete, mixed, or continuous, with a finite or infinite vector in the simplex

\[
\mathcal{P} \overset{\text{def}}{=} \{(p_1, p_2, \ldots) : p_i \geq 0 \text{ and } \sum_{i=1}^\infty p_i \leq 1\}.
\]

If \(P \overset{\text{def}}{=} (p_1, p_2, \ldots) \in \mathcal{P}\) is a distribution, then its discrete probability

\[
p \overset{\text{def}}{=} \sum_i p_i
\]

is the total probability of all discrete elements, its continuous probability

\[
q \overset{\text{def}}{=} 1 - p
\]

is the probability of all continuous elements, its discrete support

\[
K \overset{\text{def}}{=} \{i : p_i > 0\}
\]

is the set of discrete positive-probability elements, its discrete size is

\[
k \overset{\text{def}}{=} |K|,
\]

the cardinality of the discrete support, and its total size

\[
t \overset{\text{def}}{=} \begin{cases}
    k & \text{if } P \text{ is discrete} \\
    \infty & \text{if } P \text{ is mixed or continuous}
\end{cases}
\]

is the cardinality of the set of all positive-probability elements, discrete or continuous. Note that if \(P\) is continuous then \(k = 0\), and that otherwise \(k\) is positive and can be finite or infinite.

Since the order of the probabilities \(p_i\) does not affect the induced pattern probability, we typically assume that \(p_i \geq p_{i+1}\), namely, \(P = (p_1, p_2, \ldots)\) is monotone, and therefore lies in the monotone simplex

\[
\mathcal{P}_M \overset{\text{def}}{=} \{(p_1, p_2, \ldots) : p_i \geq 0, p_i \geq p_{i+1}, \text{ and } \sum_{i=1}^\infty p_i \leq 1\}.
\]

The discrete support of \(P\) is then

\[
K = [k] \overset{\text{def}}{=} \{1, 2, \ldots, k\},
\]

where \([\infty] = \{1, 2, \ldots\} \overset{\text{def}}{=} \mathbb{P}\). When describing the distribution \(P\), we can therefore omit trailing 0’s. For example, \((1)\) is the discrete distribution over a single element, \((3, 2)\) represents a mixed distribution over two elements with respective probabilities \(3\) and \(2\) and a continuous part with probability \(0.5\), and \(()\) represents a continuous distribution.
3.3 Formal definition

Following is a formal definition of pattern probabilities. It is useful for algorithmic computation of pattern probabilities, but for just an intuitive understanding it can (and might best) be skipped.

Let $A$ and $B$ be sets. By analogy with the notation $B^A$ for functions from $A$ to $B$ and $\binom{A}{2}$ for 2-element subsets of $A$, let $B^A$ denote the collection of all injections (one-to-one functions) from $A$ to $B$. Recall that $m = m(\bar{\psi})$, denotes the number of symbols in $\bar{\psi}$. Then $f \in K^{|m|}$ maps any pattern

$$\bar{\psi} = \psi_1 \psi_2 \ldots \psi_n \in [m]^n$$

to a sequence

$$f(\bar{\psi}) \overset{\text{def}}{=} f(\psi_1)f(\psi_2)\ldots f(\psi_n) \in K^n$$

with pattern $\bar{\psi}$.

As before, let $\mu_j$ denote the number of times $j$ appears in $\bar{\psi}$. Since we assume that $P$ is monotone, the probability of $\bar{\psi}$ can be expressed as

$$P(\bar{\psi}) = \sum_{f \in \{i: \mu_i = 1\}} q|I| \sum_{f \in [k]|m|-I} \prod_{j \in [m]-I} p_{f(j)}^{\mu_j}.$$  

Note that the number of products depends on the discrete size $k$, but even if $k = \infty$, the sum converges to a unique value. If the pattern is canonical, the equation can be rewritten as

$$P(\bar{\psi}) = \sum_{i = \max(0, m - k)}^{\varphi_1} \left( \begin{array}{c} \varphi_1 \\ i \end{array} \right) q^i \sum_{f \in [k]|m|-I} \prod_{j = 1}^{m-i} p_{f(j)}^{\mu_j}.$$  

If the distribution is discrete, only the $i = 0$ term may be positive, and we obtain

$$P(\bar{\psi}) = \sum_{f \in [k]|m|} \prod_{j = 1}^{m} p_{f(j)}^{\mu_j}. \quad (1)$$

Sometimes a different formula for $P(\bar{\psi})$ is useful. For a set $A$ and nonnegative integers $n_1, \ldots, n_k$ summing to at most $|A|$, we let $\binom{A}{n_1, \ldots, n_k}$ denote the collection of sequences $J_1, \ldots, J_k$ of disjoint subsets of $A$ where $|J_i| = n_i$. For example,

$$\binom{[4]}{1, 2} = \left\{ \{(1), \{2, 3\}\}, \{(1), \{2, 4\}\}, \{(1), \{3, 4\}\}, \{(2), \{1, 3\}\}, \ldots, \{(4), \{2, 3\}\} \right\}.$$  

Clearly,

$$\left| \binom{A}{n_1, \ldots, n_k} \right| = \frac{|A|!}{n_1! \cdot \ldots \cdot n_k! \cdot (|A| - n_1 - \ldots - n_k)!}.$$  

The general pattern-probability formula can then be written in terms of the prevalences,

$$P(\bar{\psi}) = \sum_{i = \max(0, m - k)}^{\varphi_1} \left( \begin{array}{c} \varphi_1 \\ i \end{array} \right) q^i \sum_{J_1, \ldots, J_{\mu_{\max}} \in \binom{[k]}{\mu_{\max}}} \prod_{j = 1}^{\mu_{\max}} \mu_j ! \prod_{j \in J_\mu} p_{\mu}^{\mu_j}.$$  

As can be seen from this equation, for any probability $P$, the induced pattern probability $P(\bar{\psi})$ depends only on the pattern’s profile $\bar{\varphi}(\bar{\psi})$. For example, $P(\bar{\Psi} = 112) = P(\bar{\Psi} = 121) = P(\bar{\Psi} = 122)$.  

11
3.4 Pattern maximum likelihood

Finally, the pattern and profile maximum likelihood (PML) probability of a pattern \( \bar{\psi} \) is

\[
\hat{P}(\bar{\psi}) \overset{\text{def}}{=} \sup_{P \in \mathcal{P}_M} P(\bar{\psi}),
\]

the highest probability assigned to \( \bar{\psi} \) by any distribution, which by symmetry we assume lies in the monotone simplex. One of the first results we prove is that \( \hat{P}(\bar{\psi}) \) is always achieved by some distribution \( \hat{P} = \hat{P}_{\bar{\psi}} = (p_1, p_2, \ldots) \), which we call the PML distribution of \( \bar{\psi} \). We denote its total size by \( \hat{t}_{\bar{\psi}} \), its discrete size by \( \hat{k}_{\bar{\psi}} \), and its continuous probability by \( \hat{q}_{\bar{\psi}} \). Since as noted earlier, a pattern’s probability is determined by its profile, so do the PML probability and distribution, and in the above notation, the pattern can be replaced by its profile.

This terminology should be interpreted with some caution. The probability induced on the empty pattern and the pattern 1, is 1 for all underlying distributions, hence we call these two patterns and their profiles, trivial, and all other patterns and profiles nontrivial. While all nontrivial patterns we have encountered have a unique PML distribution, we do not know whether this holds always. Therefore all statements made in the paper apply to all possible PML distributions. For example, when we say that the PML support of a pattern is finite, that means that the support of every PML distribution, whether the pattern has one or more, is finite.

As can be observed from the informal or formal probability definitions, \( P(\bar{\psi}) \) is a symmetric function of \((p_1, p_2, \ldots)\) and for finite \( k \) it is a degree-\( n \) single-orbit symmetric polynomial. The problem of determining the PML probability and distribution is therefore that of maximizing this polynomial over all distributions \( P \) in the monotone simplex \( \mathcal{P}_M \).

Some PML distributions are easily found. Constant sequences, where a single symbol repeats, have pattern \( 1 \ldots 1 = 1^n \) and profile \( n \). Since any distribution over a single-element set always yields the constant sequence, \( \hat{P}_{n^1} = (1) \), implying that \( \hat{k}_{n^1} = t_{n^1} = 1 \), \( \hat{q}_{n^1} = 0 \), and \( \hat{P}(\bar{\Psi} = 1^n) = 1 \).

All-distinct sequences, where every symbol appears once, have pattern \( 12 \ldots n \) and profile \( 1^n \). If \( P \) is a continuous distribution over an interval, then the probability that an i.i.d. sequence drawn according to \( P \) consists of all distinct symbols is 1. Hence the high-profile distribution is the continuous distribution \( \hat{P}_{1^n} = () \), implying that \( \hat{k}_{1^n} = 0 \), \( t_{1^n} = \infty \), \( \hat{q}_{1^n} = 1 \), and \( \hat{P}(\bar{\Psi} = 12 \ldots n) = 1 \).

Finally, in the introduction we saw that among all discrete distributions, \( \{\frac{1}{2}, \frac{1}{2}\} \) maximizes the probability of the pattern 112. We now show that the statement holds for all distributions, not only discrete. Since any repeating value must derive from the distribution’s discrete part, the probability that any distribution \((p_1, p_2, \ldots)\), where now the probabilities are at most one, assigns to the 112 pattern is

\[
\sum_i p_i^2 (1 - p_i) = \sum_i p_i \cdot [p_i (1 - p_i)] \leq \sum_i p_i \cdot \frac{1}{4} \leq \frac{1}{4}.
\]

Unlike these simple examples, establishing the high-profile distributions of all other profiles, or some of their properties, seems hard and is the subject of this and the subsequent papers.
4 Results

4.1 This paper

As observed earlier, the PML probability of some patterns, such as 12...n, can be arbitrarily approached by discrete distributions, but is not achieved by any of them. In Section 6 we show that when mixed distributions are allowed, the PML probability can always be achieved. Namely, that for any pattern, the supremum of the pattern probabilities is attained by some distribution in $\mathcal{P}_M$. In addition to its own interest, this result is also useful for proving properties of PML distributions later on. The techniques used to prove this result are analytical and differ from the combinatorial considerations utilized in the rest of the paper.

??? Sufficient statistics ???

To evaluate the efficacy of the PML approach, we study the convergence of the PML distribution to the one underlying the data sample. In Section 8, we show that under the $\ell_\infty$ norm, the PML distribution is uniformly $\frac{3}{ln n}$-consistent over $\mathcal{P}_D$.

A more stringent distance measure is the $\ell_1$ norm, but in Section 10 we show that no estimator is uniformly consistent over $\mathcal{P}_D$ under the this norm. Yet we show, that while not uniformly so, the PML distribution is still $\ell_1$ consistent over $\mathcal{P}_D$. The proof is provided in the Appendix.

Finally in Section 11, we show the the high-profile distribution can be viewed as a smoothed version of the empirical-frequency estimator.

4.2 Next paper

The properties of high-profile distributions can be broadly divided into two categories. In Section 14, we consider several properties of high-profile distributions: their total support size, discrete size, continuous probability, and an ordering of patterns based on their high-profile probabilities. These results do not assume any specific structure to the pattern in question. Deriving the high-profile distribution of profiles in general, appears to be a hard problem, and therefore in Section 15 we consider some simple profiles, whose structure makes the problem somewhat tractable. We describe the results in each of these sections in greater detail below.

The existence and probability of the continuous part of the high-profile distribution are considered in Section 14.1. Clearly, if $\varphi_1 = 0$, namely every symbol appears at least twice, then $\hat{P}$ is discrete, namely has no continuous part. We show that this holds also when $\varphi_1 = 1$, namely, exactly one of the symbols appears once. A more refined question concerns the continuous probability $\hat{q}$. In Theorem 19, we provide a simple bound on the continuous probability, showing that for all nontrivial profiles,

$$\hat{q} \leq \frac{\varphi_1}{n},$$

namely, the continuous probability of the high-profile distribution is at most the fraction of elements appearing once.

The example of ten elements, each appearing twice, showed that the discrete support of the high-profile distribution can be larger than the number of symbols observed in the data. In subsections 14.2 to 14.5 we evaluate its possible sizes.

We begin in subsection 14.2 by showing that the discrete size of the high-profile distribution is always finite. This result is also useful as it allows us later to assume a concrete (finite) form of the high-profile distribution.
In subsection 14.4 we refine the finitude result by upper bounding the size of the high-profile support. Recall that \( m \) is the number of symbols, and that \( \mu_{\text{min}} \) is the smallest number of times any symbol appears in a pattern. In Theorem 26 we show that for all nontrivial profiles,

\[
\hat{t} \leq m + \frac{m - 1}{2^{\mu_{\text{min}} - 2}},
\]

(2)

implying also the corresponding bound on \( \hat{k} \). In particular, if

\[
\mu_{\text{min}} > \log(m + 1),
\]

\[
\hat{k} = \hat{t} = m,
\]

(3)

namely, if the number of times each symbol appears is at least the logarithm of the number of symbols, then the high-profile support does not use any additional symbols.

In subsection 14.5 we turn to lower bounds on \( \hat{t} \). Recall that \( \mu_{\text{max}} \) is the maximum number of times any symbol appears. In Theorem 30 we show that

\[
\hat{t} \geq m - 1 + \left( \frac{\sum_{\mu} \mu \cdot 2^{-\mu} - 2^{-\mu_{\text{max}}}}{2^{\mu_{\text{max}} - 2}} \right).
\]

(4)

In particular, complementing (23), if

\[
\mu_{\text{max}} < \log \left( \sqrt{m} + 1 \right),
\]

namely, the number of times each symbol appears is at most half the logarithm of the number of symbols, then

\[
\hat{t} > m,
\]

namely, the high-profile support is larger than the number of symbols observed.

In Subsections 15.1 to 15.4, we use and extend the results hitherto obtained to determine the high-profile distribution of some short and simple profiles.

In Subsection 15.1, we determine \( \hat{P} \) for all profiles of length at most 4. For example, we show that the high-profile distribution of the profile \( 2^11^1 \), corresponding to the sequence @@# mentioned in the introduction, is indeed \( (1/2, 1/2) \). Perhaps the most “interesting” of these profiles is \( 2^11^2 \), where one symbol appears twice and two symbols appear once, as in the sequence @#$@. This is the shortest profile for which the high-profile discrete support size—5, is larger than the number of symbols appearing—3. Table 2 summarizes the results proven in this subsection.

In Subsection 15.2, we consider profiles of the form \( \bar{\varphi} = r^11^u \), namely, one symbol appears \( r \) times and \( u \) symbols appear once each. We show that for certain values of \( r, u \), \( \hat{P}_{\bar{\varphi}} = \left( \frac{r}{r+u} \right) \), a mixed distribution comprising a single symbol with probability \( \frac{r}{r+u} \) and a continuous part with probability \( \frac{u}{r+u} \).

In Subsection 15.3, we consider the high-profile distribution of the quasi-uniform profile where each of \( m \) symbols repeats \( r \) or \( r+1 \) times, i.e., profiles of the form \( \bar{\varphi} = (r+1)^{m-t}r^t \). We show that \( \hat{P} \) is uniform and distributed over

\[
\hat{k} = \min \left\{ k \geq m : \left( 1 + \frac{1}{k} \right)^{rm+t} \left( 1 - \frac{m}{k+1} \right) > 1 \right\}.
\]
Table 1: High-profile distribution of profiles of length at most 4.

<table>
<thead>
<tr>
<th>$\varphi$</th>
<th>Canonical $\psi$</th>
<th>$\hat{P}_\varphi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^1$</td>
<td>1</td>
<td>any distribution</td>
</tr>
<tr>
<td>$2^1, 3^1, 4^1$</td>
<td>11, 111, 1111</td>
<td>(1)</td>
</tr>
<tr>
<td>$1^2, 1^3, 1^4$</td>
<td>12, 123, 1234</td>
<td>()</td>
</tr>
<tr>
<td>$2^11^1$</td>
<td>112</td>
<td>(1/2, 1/2)</td>
</tr>
<tr>
<td>$3^11^1$</td>
<td>1112</td>
<td>(1/2, 1/2)</td>
</tr>
<tr>
<td>$2^2$</td>
<td>1122</td>
<td>(1/2, 1/2)</td>
</tr>
<tr>
<td>$2^11^2$</td>
<td>1123</td>
<td>(1/5, 1/5, 1/5, 1/5, 1/5)</td>
</tr>
</tbody>
</table>

Figure 5: High-profile distribution of binary sequences for increasing data lengths. X axis: Fraction of 0s. Y axis: $p$ where the High-Profile distribution is $(p, 1 - p)$. Dotted line: $n = 10$, dashed line: $n = 100$, solid line $n = \infty$.

symbols. In particular, as mentioned in the introduction, $\hat{k} = 12$, when $m = 10, r = 2$, and $t = 0$. We show that follows that for fixed $r$, as $m$ grows, $\hat{k}/m$ approaches the solution of

$$x \log \frac{x}{x - 1} = r.$$  

For example, for $r = 2$, the ratio $\hat{k}/m$ approaches 1.258.., hence the discrete support is a constant factor larger than the number of symbols observed. We also show that if $r = 1$ and $t \geq 1$ is a constant, then

$$\lim_{m \to \infty} \frac{t(r+1)^{(m)}r^{m-t(m)}}{m^2} = \frac{1}{2t}.$$  

In subsection 15.4 we consider sequences consisting of just two symbols. Let $n_0$ and $n_1$, where wolog $1 \leq n_0 \leq n_1 \leq n - 1$, denote the number of times each of the two symbols appears in a sequence of length $n$. Combining the upper bound (22) with a result by Alon [?], we show in Theorem 49 that if $(n_1 - n_0)^2 \leq n$, then $\hat{P} = (0.5, 0.5)$, and if $(n_1 - n_0)^2 > n$, then $\hat{P} = (\frac{1}{1+\alpha}, \frac{\alpha}{1+\alpha})$, where $\alpha$ is the unique root in $(0, 1)$ of the polynomial

$$n_0 \cdot x^{n_1-n_0+1} - n_1 \cdot x^{n_1-n_0} + n_1 \cdot x - n_0.$$  

Figure 13 plots the probability $p = \alpha/(\alpha + 1)$ indicated by the high-profile distribution as a function of $n_0/n$ for various sequence lengths $n$. Observe that $p$ is always between $n_0/n$ and $\frac{1}{2}$.

As Figure 13 shows, while the high-profile and maximum-likelihood distributions of binary sequences differ for every block length $n$, as $n$ tends to infinity, they converge to each other.

5 Sufficient statistics

We will show that for certain a class of estimation problems the pattern is a sufficient statistic. For $s \in \mathbb{P}$, let $\Omega_s$ be a distribution on all ordered subsets of $\mathbb{P}$ of size $s$. The distribution $\Omega_s$ is
isotropic if for any permutation \( \pi : [s] \xrightarrow{1-1} [s] \), and all \( \{a_1, a_2, \ldots, a_s\} \subset \mathbb{P} \),
\[
\Omega_k(a_1, a_2, \ldots, a_s) = \Omega_k(\pi(1), \pi(2), \ldots, \pi(s)).
\] (5)

As implied in Section 3.4, the monotone simplex \( \mathcal{P}_d \) can be viewed as representing the collection of all probability multisets. Let \( \mathcal{P}_p(k) \) denote the subset of \( \mathcal{P}_d \) that consists only of multisets of size \( k \).

In our estimation problem, let \( P = (p_1, p_2, \ldots, p_k) \in \mathcal{P}_p(k) \) be the probability multiset underlying the observed data. We will now describe how the data is generated from this multiset. A distribution \( \tilde{P} \) is selected at random by choosing an ordered subset \((A_1, A_2, \ldots, A_k)\) according to an isotropic distribution \( \Omega_k \) and then setting
\[
\tilde{P}(a) = \begin{cases} p_i & \text{if } a = A_i, \\ 0 & \text{otherwise}. \end{cases}
\]

A sequence \( X^n = X_1, X_2, \ldots, X_n \) of i.i.d. random variables is then generated according to \( \tilde{P} \). The probability distribution of \( X^n \) is therefore
\[
Pr(X^n) = \sum_{\mathcal{A}(X^n) \subseteq \mathcal{A} \subseteq \mathcal{P}, \pi: \mathcal{A} \xrightarrow{1-1} [k]} \Omega_k(\pi^{-1}(1), \pi^{-1}(2), \ldots, \pi^{-1}(k)) \prod_{i=1}^{n} p_{\pi(X_i)}
\] (6)

where \( \pi \) is a one-to-one function. We will show that for the purpose of estimating the underlying multiset \( P \), the pattern \( \psi(X^n) \) is a sufficient statistic. Note that the isotropic property of \( \Omega_k \) is chosen as a mechanism to design a setting where the estimator has no prior knowledge of the support of the underlying distribution.

**Theorem 1** For all finite \( k \), and all isotropic \( \Omega_k \), the pattern \( \psi(X^n) \) is a sufficient statistic for \( P \in \mathcal{P}_p(k) \).

**Proof** Let \( \mathcal{A}(X^n) \) denote the set of distinct symbols in the sequence \( X^n \). We will show that the probability of a sequence \( X^n \) with \( m \) distinct symbols can be written as
\[
Pr(X^n) = P(\psi(X^n))G(k, \Omega_k, \mathcal{A}(X^n))
\] (7)

where \( P(\psi(X^n)) \), the probability of the pattern depends only on the pattern and \( P \), and the function \( G(\cdot) \) does not depend on \( P \). This proves that the conditional distribution \( Pr(X^n|\psi(X^n)) \) of \( X^n \) conditioned on the pattern does not depend on \( P \) and therefore \( \psi(X^n) \) is a sufficient statistic for \( P \in \mathcal{P}_p(k) \).

Let \( P = (p_1, p_2, \ldots, p_k) \in \mathcal{P}_p(k) \) be the underlying probability multiset. Combining (6) and (5)
\[
Pr(X^n) = \sum_{\mathcal{A}(X^n) \subseteq \mathcal{A} \subseteq \mathcal{P}, \pi: \mathcal{A} \xrightarrow{1-1} [k]} \Omega_k(\pi^{-1}(1), \pi^{-1}(2), \ldots, \pi^{-1}(k)) \prod_{i=1}^{n} p_{\pi(X_i)}
\]
\[
= \sum_{\mathcal{A}(X^n) \subseteq \mathcal{A} \subseteq \mathcal{P}, \pi: \mathcal{A} \xrightarrow{1-1} [k]} \Omega_k(a_1, a_2, \ldots, a_k) \prod_{i=1}^{n} p_{\pi(X_i)}
\]
\[
= \sum_{\mathcal{A}(X^n) \subseteq \mathcal{A} \subseteq \mathcal{P}, \pi: \mathcal{A} \xrightarrow{1-1} [k]} \Omega_k(a_1, a_2, \ldots, a_k) \prod_{\pi:A \xrightarrow{1-1}[k]}^{n} p_{\pi(X_i)}
\] (8)
where \( A = \{a_1, a_2, \ldots, a_k\} \). Observing that each bijective function from \( A \) to \([k]\) can be viewed as a combination of an injection \( f \) from \( A(X^n) \) to \([k]\) and a bijection from \( A/A(X^n) \) to \([k]/f(A(X^n))\), we obtain that

\[
\sum_{\pi: A \rightarrow [k]} \prod_{i=1}^{n} p_{\pi(X_i)} = \sum_{f: [k]/A(X^n)} \prod_{i=1}^{n} p_{f(X_i)} \sum_{g: A/A(X^n) \rightarrow [k]/f(A(X^n))} 1 = P(\psi(X^n))(k - m)!
\]

where recall that \( P(\psi(X^n)) \) is probability of the pattern of \( X^n \) and it depends only on the multiset \( P \), and \( m = |A(X^n)| \) is the number of distinct symbols in \( X^n \). Equation (9) follows from the definition of the probability of a pattern for discrete distributions (see (1)). Substituting (9) in (8) we obtain that

\[
Pr(X^n) = P(\psi(X^n))(k - m)! \sum_{A(X^n) \subseteq A \subseteq \mathcal{P}, |A| = k} \Omega_k(a_1, a_2, \ldots, a_k)
\]

which is of the same form as (7).

6 Existence

Recall that the PML probability of some patterns such as 12\ldots n can be arbitrarily approached by discrete distributions, but is not achieved by any of them. We show that the PML probability of any pattern is achieved by some, possibly mixed, distribution, hence the PML multiset always exists. The existence of the PML distribution is interesting on its own and is also useful in proving some of the PML properties.

To prove that the PML probability is always attained, we view probability distributions as vectors \( P = (p_1, p_2, \ldots) \) in the monotone simplex \( \mathcal{P}_M \) and show that under the \( \ell_2 \) norm

\[
|| (p_1, p_2, \ldots) ||_2 \overset{\text{def}}{=} \left( \sum_{i=1}^{\infty} p_i^2 \right)^{1/2},
\]

\( \mathcal{P}_M \) is compact and for all patterns \( \tilde{\psi} \), \( P(\tilde{\psi}) \) is a continuous function of \( P \). We begin with the continuity proof as it is more ‘compact’, and related to pattern probabilities.

Note first that had we considered only distributions of a bounded finite size, continuity of \( P(\tilde{\psi}) \) would have been clear. The continuous probability \( q = 1 - \sum p_i \) would have been continuous in \( P \) and hence \( P(\tilde{\psi}) \) would have been a polynomial in finitely many continuous function of \( P \) hence continuous too.

However, distributions in \( \mathcal{P}_M \) may have an unbounded, even infinite number of vaules, and not even the continuity of \( q \) in \( P \) can be assured. Consider for example the sequence \( P_1, P_2, \ldots \) of distributions where

\[
P_k = \left( \begin{array}{cccc} 1 & 1 & \cdots & 1 \\ \frac{1}{k} & \frac{1}{k} & \cdots & \frac{1}{k} \\ \frac{1}{k} & \cdots & \frac{1}{k} \\ \frac{1}{k} & \frac{1}{k} & \cdots & \frac{1}{k} \end{array} \right).
\]

The \( \ell_2 \) distance between each \( P_k \) and the continuous distribution (\( \) is \( || P_k - (\) ||_2 = 1/\sqrt{k} \), hence the \( P_k \)’s converge to (\). However, since each \( P_k \) is discrete, its continuous probability is 0, and yet the limit distribution (\) is continuous, hence its continuous probability is 1.
Yet despite the discontinuity of $q$ in $P$, and a related concern stemming from the potential of infinitely many $p_i$’s, the pattern probability $P(\tilde{\psi})$ is continuous in $P$. Consider for example the pattern 11. Its induced probability under $P_k$ is $\frac{1}{2}$ which converges to 0, the same probability induced by the limit distribution (). We first show that similarly, $P(\psi)$ is continuous for all constant patterns $1^n$.

**Lemma 2** $P(1^n)$ is continuous in $P \in \mathcal{P}$ for all $n$.

**Proof** If $n = 1$ then $P(\tilde{\psi}) = P(1) = 1$ for all $P \in \mathcal{P}$, hence $P(\tilde{\psi})$ is continuous in $P$. For $n \geq 2$ we show that for all $\epsilon > 0$ there is a $\delta > 0$ such that

$$||P - Q||_2 < \delta \implies |P(1^n) - Q(1^n)| < \epsilon.$$  

Generalizing

$$p^2 - q^2 = (p + q)(p - q),$$

for all $n$,

$$p^n - q^n = (p^{n-1} + p^{n-2}q + \ldots + pq^{n-2} + q^{n-1})(p - q).$$

Hence for $p, q \in [0, 1]$,

$$|p^n - q^n| = (p^{n-1} + p^{n-2}q + \ldots + pq^{n-2} + q^{n-1}) \cdot |p - q| \leq \frac{n}{2} |p + q| |p - q|.$$  

It follows that if

$$||P - Q||_2 < \delta \overset{\text{def}}{=} \frac{\epsilon}{n},$$

then

$$|P(1^n) - Q(1^n)| = \left| \sum_{i=1}^{\infty} p_i^n - \sum_{i=1}^{\infty} q_i^n \right|$$

$$\leq \sum_{i=1}^{\infty} |p_i^n - q_i^n|$$

$$\leq \sum_{i=1}^{\infty} \frac{n}{2} (p_i + q_i) |p_i - q_i|$$

$$\leq \sum_{i=1}^{\infty} \frac{n}{2} (p_i + q_i) ||P - Q||_2$$

$$< \epsilon. \quad \Box$$

**Theorem 3** $P(\tilde{\psi})$ is continuous in $P \in \mathcal{P}$ for all patterns $\tilde{\psi}$.

**Proof** We prove by induction on $m$ that $P(\tilde{\psi})$ is continuous for all patterns with at most $m$ symbols.

For $m = 1$, $\tilde{\psi} = 1^n$ and the result was proven in the last lemma. To prove the induction step, assume that $P$ is continuous for all patterns with at most $m$ symbols, and show that it is continuous for all patterns $\tilde{\psi}$ with at most $m + 1$ symbols. Without loss of generality, $\tilde{\psi}$ is canonical, hence

$$P(\tilde{\psi}) = P(1^{\mu_1} 2^{\mu_2} \ldots m^{\mu_m} (m + 1)^{\mu_{m+1}})$$

$$= P(\{1^{\mu_1} 2^{\mu_2} \ldots m^{\mu_m} i^{\mu_{m+1}} : 1 \leq i \leq m + 1\}) - \sum_{i=1}^{m} P(1^{\mu_1} 2^{\mu_2} \ldots m^{\mu_m} i^{\mu_{m+1}})$$

$$= P(1^{\mu_1} 2^{\mu_2} \ldots m^{\mu_m}) \cdot P(1^{\mu_{m+1}}) - \sum_{i=1}^{m} P(1^{\mu_1} 2^{\mu_2} \ldots m^{\mu_m} i^{\mu_{m+1}}),$$
where the last equality follows since the distribution is i.i.d., hence, letting \( n = \mu_1 + \ldots + \mu_m \),

\[
\begin{align*}
P\{1^{\mu_1}2^{\mu_2} \ldots m^{\mu_m}(\mu_{m+1}: 1 \leq i \leq m+1) \\
&= P\left(\psi(X^n) = 1^{\mu_1}2^{\mu_2} \ldots m^{\mu_m} \& \psi(X_{n+1}^{\mu_{m+1}}) = 1^{\mu_{m+1}}\right) \\
&= P\left(\psi(X^n) = 1^{\mu_1}2^{\mu_2} \ldots m^{\mu_m}\right) \cdot P\left(\psi(X_{n+1}^{\mu_{m+1}}) = 1^{\mu_{m+1}}\right) \\
&= P\left(\psi(X^n) = 1^{\mu_1}2^{\mu_2} \ldots m^{\mu_m}\right) \cdot P(1^{\mu_{m+1}}).
\end{align*}
\]

It follows that \( P(\tilde{\psi}) \) is a sum of a finite number of functions that by induction hypothesis are continuous in \( P \), hence is continuous in \( P \).

We have therefore shown that \( P(\tilde{\psi}) \) is continuous in \( P \) for all patterns \( \tilde{\psi} \), and we now show that \( \mathcal{P}_M \) is compact.

Let \( \mathcal{X} \) be a set with a metric defined by a norm \( \| \cdot \| \). The set \( \{ x \in \mathcal{X} : \| x - x_0 \| < \epsilon \} \) is the \( \epsilon \)-ball centered at \( x_0 \in \mathcal{X} \). \( \mathcal{X} \) is complete if every Cauchy sequence in \( \mathcal{X} \) converges to some element in \( \mathcal{X} \), and it is totally bounded if for all \( \epsilon > 0 \), it is covered by finitely many \( \epsilon \)-balls. It is well known, e.g., Theorem 3.1 in [?], that \( \mathcal{X} \) is compact if and only if it is complete and totally bounded.

The next lemma shows that \( \mathcal{P}_M \) is complete, and Lemma 5 shows that \( \mathcal{P}_M \), unlike the non-monotone simplex \( P \), is totally bounded. Together, the lemmas imply that \( \mathcal{P}_M \) is compact. Combined with the continuity of \( P \), Theorem 6 concludes that the supremum of \( P(\tilde{\psi}) \) in \( \mathcal{P}_M \) is achieved.

**Lemma 4** \( \mathcal{P}_M \) is complete.

**Proof** We show that any Cauchy sequence \( P^\infty = P_1, P_2, \ldots \) of distributions in \( \mathcal{P}_M \) converges to a distribution in \( \mathcal{P}_M \).

The space \( \ell_2 \) of sequences with finite \( \ell_2 \) norm is complete, e.g., [?]. Since any \( P = (p_1, p_2, \ldots) \in \mathcal{P}_M \) consists of non-negative numbers that sum to at most one, \( P \in \ell_2 \), and hence \( \mathcal{P}_M \subseteq \ell_2 \). It follows that \( P^\infty \) converges to a distribution \( Q = (q_1, q_2, \ldots) \in \ell_2 \). All we need to show is that \( Q \in \mathcal{P}_M \), namely, \( q_j \geq q_{j+1} \geq 0 \) for all \( j \), and \( \sum_{j=1}^{\infty} q_j \leq 1 \). We prove both claims by contradiction.

Suppose that \( q_j < 0 \) for some \( j \). Then for all \( i \),

\[
\|P_i - Q\|_2 \geq |q_j|,
\]

and \( P^\infty \) cannot converge to \( Q \). A similar argument shows that \( q_j \geq q_{j+1} \) for all \( j \).

To prove \( \sum_{j=1}^{\infty} q_j \leq 1 \), suppose that

\[
\sum_{j=1}^{\infty} q_j > 1.
\]

Then for some \( t \),

\[
a \triangleq \sum_{j=1}^{t} q_j > 1.
\]

Since \( P^\infty \) converges to \( Q \), for some \( i \),

\[
\|P_i - Q\|_2 \leq \frac{a - 1}{2t},
\]

19
letting $P_i = (p_{i,1}, p_{i,2}, \ldots)$, we obtain
\[
\sum_{j=1}^{t} (p_{i,j} - q_j)^2 \leq \left( \frac{a - 1}{2t} \right)^2.
\]
Therefore, for all $j \leq t$,
\[
|p_{i,j} - q_j| \leq \frac{a - 1}{2t},
\]
and
\[
\sum_{j=1}^{t} |p_{i,j} - q_j| \leq \frac{a - 1}{2}.
\]
Hence,
\[
\sum_{j=1}^{\infty} p_{i,j} \geq \sum_{j=1}^{t} p_{i,j} \geq \sum_{j=1}^{t} q_j - \sum_{j=1}^{t} |p_{i,j} - q_j| \geq a - \frac{a - 1}{2} = \frac{a + 1}{2} > 1.
\]
where the second inequality follows from the triangle inequality, contradicting the assumption that, since $P_i \in \mathcal{P}_M$, $\sum_{j=1}^{\infty} p_{i,j} \leq 1$. Therefore $\sum_{j=1}^{\infty} q_j \leq 1$, and $Q \in \mathcal{P}_M$.

Observe that the non-monotone $\mathcal{P}$ is not totally bounded. Every point in $\mathcal{P}$ exceeds $\frac{1}{2}$ in at most one coordinate. Hence the coordinate-wise maximum of every finite collection of points in $\mathcal{P}$ exceeds $\frac{1}{2}$ in finitely many coordinates. It follows that for any finite collection of $\frac{1}{2}$-balls centered at points in $\mathcal{P}$, there is a coordinate where each of the balls is centered at or below $\frac{1}{2}$, and thus does not cover the point $(0, \ldots, 0, 1, 0, \ldots) \in \mathcal{P}$. Yet while the non monotone simplex $\mathcal{P}$ is not totally bounded, and hence not compact, the monotone simplex $\mathcal{P}_M$ is totally bounded and hence compact.

**Lemma 5** \( \mathcal{P}_M \) is totally bounded.

**Proof** If $P = (p_1, p_2, \ldots) \in \mathcal{P}_M$, then $\sum_{i=1}^{\infty} p_i \leq 1$, and for all $i$, $p_i \geq p_{i+1}$, hence $p_i \leq 1/i$ for all $i$. Given $\epsilon > 0$, let $t$ be such that
\[
\sum_{i=t+1}^{\infty} \frac{1}{i^2} < \frac{\epsilon^2}{2}.
\]
Then the tail probability of any $P \in \mathcal{P}_M$ satisfies
\[
\sum_{i=t+1}^{\infty} p_i^2 < \frac{\epsilon^2}{2}.
\]

Let $\mathcal{P}_{M,t}$ be the restriction of $\mathcal{P}_M$ to the first $t$ coordinates. $\mathcal{P}_{M,t}$ is totally bounded for any $t$. For example, given $\epsilon > 0$, take all balls whose centers are in $\mathcal{P}_{M,t}$ with coordinates that are integer multiples of $\sqrt{\epsilon/t}$. Hence $\mathcal{P}_{M,t}$ can be covered by some finite number, say $M$, of $\epsilon/\sqrt{2}$-balls centered at $P_1, \ldots, P_M \in \mathcal{P}_{M,t}$.

For $i = 1, 2, \ldots, M$, construct $Q_i$ by appending infinitely many zeros to $P_i$. For example, if $P_i = (p_{i,1}, \ldots, p_{i,t})$, then $Q_i = (q_{i,1}, q_{i,2}, \ldots) = (p_{i,1}, \ldots, p_{i,t}, 0, 0, \ldots)$. We show that the $\epsilon$-balls around $Q_1, \ldots, Q_M$ cover $\mathcal{P}_M$.

For any $P = (p_1, p_2, \ldots) \in \mathcal{P}_M$, there exists an $i$ such that
\[
\sum_{j=1}^{i} (p_j - q_{i,j})^2 = \sum_{j=1}^{i} (p_j - p_{i,j})^2 < \frac{\epsilon^2}{2},
\]
which, by the choice of \( t \), implies

\[
\sum_{j=1}^{\infty} (p_j - q_{i,j})^2 = \sum_{j=1}^{t} (p_j - q_{i,j})^2 + \sum_{j=t+1}^{\infty} (p_j - 0)^2 < \frac{\epsilon^2}{2} + \frac{\epsilon^2}{2} = \epsilon^2.
\]

Therefore \( Q_1, \ldots, Q_M \) cover \( \mathcal{P}_M \). \( \square \)

We end this section by concluding that the supremum of \( P(\tilde{\psi}) \) over all \( P \in \mathcal{P} \) is achieved.

**Theorem 6** For all \( \tilde{\psi} \), there exists \( P \in \mathcal{P}_M \) such that

\[
P(\tilde{\psi}) = \hat{P}(\tilde{\psi}).
\]

**Proof** From Lemmas 4 and 5, \( \mathcal{P}_M \) is compact and from Theorem 3, \( P(\tilde{\psi}) \) is continuous in \( P \) over \( \mathcal{P} \supseteq \mathcal{P}_M \). Standard arguments, e.g., [?], imply that \( \hat{P}(\tilde{\psi}) \), the supremum of \( P(\tilde{\psi}) \) over \( \mathcal{P}_M \) is attained by some distribution \( P \in \mathcal{P}_M \). \( \square \)

## 7 Consistency

subsectionThe estimation problem Recall that

\[
\mathcal{P}_D \overset{\text{def}}{=} \{ P \in \mathcal{P} : q = 0 \}
\]

is the set of all discrete distributions and

\[
\mathcal{P}_{M,D} \overset{\text{def}}{=} \{ P \in \mathcal{P}_M : q = 0 \}
\]

is the set of all discrete distributions in the monotone simplex \( \mathcal{P}_M \). Further, let \( S : \mathcal{P}_D \to \mathcal{P}_{M,D} \) denote the sorting function that maps discrete distributions to the sorted versions in \( \mathcal{P}_{M,D} \). As explained in Section 1, we are interested in statistical quantities that are determined by just the multiset of probabilities, with no regard to their associated outcomes. Therefore, given a sequence \( \overline{X} = X_1, X_2, \ldots, X_n \) drawn \( i.i.d. \) according to an unknown discrete distribution \( P \in \mathcal{P}_D \), we would like to find an estimate \( \hat{P} = f(\overline{X}) \in \mathcal{P}_M \) that is “close” to \( S(P) \). Note that our estimates may be mixed, \( i.e., \) from the set \( \mathcal{P}_M \), but the estimands are from \( \mathcal{P}_{M,D} \). We study the quality of our estimators according to two measures. The \( \ell_\infty \) distance between two distributions \( P = (p_1, p_2, \ldots) \) and \( Q = (q_1, q_2, \ldots) \) in \( \mathcal{P}_M \) is

\[
\|P - Q\|_\infty \overset{\text{def}}{=} \max_i |p_i - q_i|,
\]

and the \( \ell_1 \) or *variational* distance is

\[
\|P - Q\|_1 \overset{\text{def}}{=} \sum_{i=1}^{\infty} |p_i - q_i|,
\]

Let \( \| \cdot \| \) represent the metric of interest and let \( \overline{X} = X_1, X_2, \ldots, X_n \) denote an \( i.i.d. \) sequence drawn according to some distribution \( P \in \mathcal{P}_D \). We are seeking an estimator \( f(\cdot) \) such that \( \|f(\overline{X}) - S(P)\| \) is small with high probability. To present this notion formally, we require definitions of consistency which we present in the next subsection.
7.1 Consistency

Let $\bar{X} = X_1, X_2, \ldots, X_n$ be drawn i.i.d. according to a discrete distribution $P$ from some collection of distributions $\mathcal{P}$. Let $\{f_n(\cdot)\}_{n=1}^{\infty}$ denote a sequence of estimators for some function of $P$, say $g(P)$ where $f_n(\cdot)$ is a function of $X_1, X_2, \ldots, X_n$. Henceforth we will abbreviate such a sequence by $\{f_n\}$. For a given distance metric, the sequence $\{f_n\}$ is consistent for $P$ if for all $P \in \mathcal{P}$,

$$f_n(X_1, X_2, \ldots, X_n) \xrightarrow{P} g(P),$$

i.e., the sequence of estimates $\{f_n(X_1, X_2, \ldots, X_n)\}$ converges in probability to $g(P)$, where the convergence in probability is with respect to the distance metric in question. In other words, if $\{f_n\}$ is consistent for $g(P), P \in \mathcal{P}$, then for all $\epsilon > 0$, and all $P \in \mathcal{P}$,

$$\lim_{n \to \infty} P(||f_n(X_1, X_2, \ldots, X_n) - g(P)|| > \epsilon) = 0.$$

The sequence $\{f_n\}$ is uniformly consistent for $g(P), P \in \mathcal{P}$, if for all $\epsilon > 0$

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}} P(||f_n(X_1, X_2, \ldots, X_n) - g(P)|| > \epsilon) = 0.$$

Clearly, uniform consistency is a more stringent requirement and has the property that for any given $\epsilon > 0$, $\delta > 0$, one can decide on a sample size $n$ that would suffice for $f_n$ to return an estimate that is within distance $\epsilon$ of $g(P)$ with probability at least $1 - \delta$.

The definition of consistency can be refined to incorporate the notion of the rate of convergence of a sequence of estimators. To begin with we state the definitions of the probabilistic version of the Landau symbol, $O(\cdot)$, namely, $O_P(\cdot)$.

A sequence of random variables $\{X_n\}$ distributed according to $P$, is bounded in probability if for any $\delta > 0$ there exist $M$ and $n_0$ such that for all $n > n_0$,

$$P(||X_n|| > M) \leq \delta.$$

Let $\{k_n\}$ be a sequence of real numbers. We say that $X_n = O_P(1/k_n)$ if $k_n X_n$ is bounded in probability.

Let $\bar{X} = X_1 X_2 \ldots X_n$ be distributed i.i.d. according to some $P \in \mathcal{P}$. A sequence of estimators $\{f_n\}$ is $k_n$-consistent for $g(P), P \in \mathcal{P}$, if for every $P \in \mathcal{P}$

$$f_n(\bar{X}) - g(P) = O_P(1/k_n).$$

As in the case of consistency, the sequence of estimators $\{f_n\}$ is uniformly $k_n$-consistent for $g(P), P \in \mathcal{P}$, if for any $\delta > 0$ there exists $M$ and $n_0$ such that for all $n > n_0$

$$\sup_{P \in \mathcal{P}} P(k_n ||f_n(\bar{X}) - g(P)|| > M) \leq \delta.$$

We are interested in proving uniform $k_n$-consistency for sequences of estimators, where $k_n$ is usually of the form $\sqrt{n}, n, n/\log n$ etc. In particular, we are interested in deriving sequences $\{f_n\}$ of estimators that are $k_n$-consistent for $S(P), P \in \mathcal{P}_D$.

7.2 Estimators

We consider two estimators, SML and PML. For an $n$-element sequence $\pi = x_1 x_2 \ldots x_n$, let $\{\mu_1, \mu_2, \ldots, \mu_m\}$ be the multiset of multiplicities of the different symbols, where clearly

$$\sum_{i=1}^{m} \mu_i = n.$$
The SML estimator $\hat{N}_n$ is the sorted estimate

$$\hat{N}_n(\mathcal{X}) \overset{\text{def}}{=} \left( \frac{\mu(1)}{n}, \frac{\mu(2)}{n}, \ldots, \frac{\mu(m)}{n} \right),$$

where $\{\mu(i)\}$ are the multiplicities arranged in decreasing order, i.e., $\mu(i) \geq \mu(i+1)$ for $i = 1, 2, \ldots, m-1$.

The high-profile distribution was defined in Section 3.2. To recall, let $\bar{\psi} = \psi(\mathcal{X})$ denote the pattern of the observed sequence. Then the high-profile distribution is

$$\hat{P}(\bar{\psi}) = \sup_{P \in \mathcal{P}_M} P(\bar{\psi}),$$

where the supremum is over all distributions in $\mathcal{P}_M$, not just those that are discrete. We show in Theorem 6 that there exists a distribution that achieves this supremum and we denote this distribution as the high-profile distribution. However, in the absence of a proof of its uniqueness, the definition of an estimator as the high-profile distribution is ambiguous. We define the high-profile distribution $\hat{P}_n$ to be any distribution that achieves this supremum, where the subscript $n$ denotes the length of the observed sequence. If the high-profile distribution is not unique, all the results we state in the sequel apply to any high-profile distribution. We prove consistency results for the high-profile distribution and state the corresponding results without proof for the empirical-frequency estimator.

Note that both the empirical-frequency estimator and the high-profile distribution are functions of the profile of the observed sequence and do not depend on the actual symbols observed.

8 Uniform $\ell_\infty$ consistency

We prove that under the $\ell_\infty$ norm, SML is uniformly $f(n)$-consistent over $\mathcal{P}_M$ for all $f(n) = o(\sqrt{n})$, and use this result to show that PML is uniformly $f(n)$-consistent for all $f(n) = o(n^{1/4})$. One weakness of this proof technique is that the PML bound is based on that of SML, hence does not capture the improved convergence of PML observed in practice.

We begin with a well-known bound on the probability that the sample average will differ markedly from the expected mean.

**Theorem 7 (Chernoff bound, e.g., [?])** Let $X_1, X_2, \ldots, X_n$ be i.i.d. $B(p)$ random variables with average

$$Y \overset{\text{def}}{=} \frac{|\{i \leq n : X_i = 1\}|}{n}.$$

Then for all $\epsilon > 0$,

$$P(Y \leq p - \epsilon) \leq e^{-\frac{\epsilon^2 n}{2p}}$$

and

$$P(Y \geq p + \epsilon) \leq e^{-\frac{\epsilon^2 n}{2(p+\epsilon)}}. \quad \square$$

For i.i.d. Bernoulli $(p_0, p_1)$ distributions, $|N_0 - p_0| = |N_1 - p_1|$, and the Chernoff bounds imply that for all such distributions, uniformly over $(p_0, p_1)$,

$$P(\exists i : |N_i - p_i| \geq \epsilon) \leq 2e^{-\frac{\epsilon^2 n}{4}}.$$

We show that a similar bound holds uniformly for all i.i.d. distributions, regardless of the alphabet-size. We will use the following result.
Lemma 8  For \( \alpha \geq 2 \),
\[
\sum_{p_i : \sum p_i \leq 1} e^{-\frac{p_i}{\alpha}} = \sum_{p_i : \sum p_i \leq 1/\alpha} e^{-\frac{1}{\alpha}} \leq e^{-\alpha}.
\]

Proof  Differentiating twice,
\[
\left(e^{-\frac{1}{\alpha}}\right)^{''} = \frac{1 - 2p}{p^4 e^{1/p}}.
\]
Hence letting \( e^{-1/\alpha} \overset{\text{def}}{=} \lim_{p \to 0} e^{-1/p} = 0 \), we see that \( e^{-1/p} \) is convex for \( p \in [0, \frac{1}{2}] \). Now if \( f(0) = 0 \) and for \( a_i \geq 0 \), \( f \) is convex over \([0, \sum a_i]\), then
\[
f(a_i) \leq \frac{a_i}{\sum a_i} f(\sum a_i),
\]
and the lemma follows from
\[
\sum f(a_i) \leq \sum \frac{a_i}{\sum a_i} f(\sum a_i) = f(\sum a_i).
\]

The lemma extends the Chernoff bound to arbitrary distributions.

Lemma 9  For any \( P \in \mathcal{P}_d \) and all \( \epsilon \geq 4/\sqrt{n} \),
\[
P(\exists i : |N_i - p_i| \geq \epsilon) \leq 3e^{-\frac{2n}{\epsilon^2}}.
\]

Proof  From the union and Chernoff bounds and the previous lemma,
\[
P(\exists i : N_i \leq p_i - \epsilon) \leq \sum_{i: p_i \geq \epsilon} P(N_i \leq p_i - \epsilon) \leq \sum_i e^{-\frac{2n}{3p_i}} \leq e^{-\frac{2n}{\epsilon^2}}.
\]
In the reverse direction, we need to consider only the discrete probabilities. For \( i \) such that \( p_i > \epsilon/3 \),
\[
P(N_i > p_i + \epsilon) \leq e^{-\frac{2n}{12p_i \epsilon}} \leq e^{-\frac{2n}{3p_i}}.
\]
The remaining \( i \)'s have \( p_i \leq \epsilon/3 \) and they can be partitioned into sets \( S_1, S_2, \ldots \) such that \( \epsilon/6 < \sum_{i \in S_j} p_i \leq \epsilon/3 \), and possibly one final set \( S_0 \) with \( \sum_{i \in S_0} p_i \leq \epsilon/6 \). To see that, consider the remaining \( p_i \)'s in decreasing order and sequentially create each \( S_j \) by repeatedly adding \( p_i \)'s till the first time their sum strictly exceeds \( \epsilon/6 \), then proceeding with \( S_{j+1} \). Group the elements left at the end of the process in \( S_0 \). Let \( N_j \overset{\text{def}}{=} \sum_{i \in S_j} N_i \) and \( p_j \overset{\text{def}}{=} \sum_{i \in S_j} p_i \). For \( j \geq 1 \), \( \frac{\epsilon}{6} \leq p_j \leq \frac{\epsilon}{3} \), hence from the Chernoff bound,
\[
P(\exists i \in S_j : N_i > p_i + \epsilon) \leq P(N_j > \epsilon) \leq P(N_j > p_j + 2\epsilon/3) \leq e^{-\frac{(2\epsilon/3)^2n}{2(p_j^2 + 2p_j/3)\epsilon}} = e^{-\frac{2n}{12p_j^2/3}},
\]
and a more careful evaluation of the second inequality yields
\[
P(N_j > \epsilon) \leq P(N_j > p_j + (\epsilon - p_j)) \leq e^{-\frac{\min_{1/6 \leq \epsilon \leq 1/3} 3\epsilon^2(1-n/2p_j)^2n}{192p_j^2}} \leq e^{-\frac{2n}{192p_j^2}} \leq e^{-\frac{2n}{8p_j^2}}.
\]
For \( S_0 \),
\[
P(\exists i \in S_0 : N_i > p_i + \epsilon) \leq P(N^0 > \epsilon) \leq P(N^0 > p^0 + 5\epsilon/6) \leq e^{-\frac{25n}{12p^0}}.
\]
Hence
\[
P(\exists i : N_i > p_i + \epsilon) \leq \sum_{i : p_i > \epsilon/3} e^{-\frac{2n}{4p_i}} + \sum_{j \geq 1} e^{-\frac{2n}{8p_j^2}} + e^{-\frac{25n}{12p^0}} \leq 2e^{-\frac{2n}{8}},
\]
and the lemma follows. \( \square \)
Observe that the condition $\epsilon \geq 4/\sqrt{n}$ is not too stringent as for smaller $\epsilon$ the bound is merely a constant. The bound is most useful when it diminishes to 0, namely $\epsilon \gg 4/\sqrt{n}$, the range in which we apply it in Theorem 11.

The lemma shows that with high probability each $N_i$ is close to $p_i$. We now show that when two sequences are sorted, their $\ell_\infty$ distance decreases. Hence after sorting the $N_i$'s, the maximum difference between them and the $p_i$'s will further decrease.

For length-2 sequences $a_1, a_2$ and $b_1, b_2$, the $\ell_\infty$ decrease is intuitively clear. If both sequences increase, or both decrease, sorting will not change the $\ell_\infty$ distance between them. If one sequence increases and the other decreases, sorting will transpose one of the sequences, and the following simple argument shows that the maximum difference between the elements will decrease. Note that if $x' + y' = x + y$ and $|x' - y'| \leq |x - y|$, then $\max(|x'|, |y'|) \leq \max(|x|, |y|)$. Since $a_1 - a_2$ and $b_1 - b_2$ have opposing signs,

$$\max(|a_1 - b_2|, |a_2 - b_1|) \leq \max(|a_1 - b_1|, |a_2 - b_2|).$$

A similar argument works for all finite-length sequences, say $a^n, b^n$. First sort one sequence, say $a^n$, while applying the same permutation to the other. This will leave all the absolute differences $|a_i - b_i|$, and hence their maximum, unchanged. Then sort the sequence $b^n$ via a finite number of transpositions, each exchanging the values of some unsorted $b_i$ and $b_j$. The argument above shows that the maximum absolute difference decreases at each step.

Sorting infinite-length sequences may require infinitely many transpositions, but as we now show, the resulting $\ell_\infty$ distance is still smaller. While the $N_i$ sequences considered in this section are finite, this generalization may prove useful for other estimators.

A finite or infinite sequence $\overline{a} = a_1, a_2, \ldots$ is sortable if there is a permutation $\sigma$ of the indices such that the resulting sequence $\overline{a'} \stackrel{\text{def}}{=} a_{\sigma(1)}, a_{\sigma(2)}, \ldots$ is non-increasing. All finite sequences are clearly sortable. The infinite sequences $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ as well as $-2, -1, -4, -3, -6, -5, \ldots$ and $\frac{1}{1}, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \frac{1}{9}, \frac{1}{11}, \frac{1}{13}, \ldots$ are sortable, while $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots$, as well as $1, 2, 1, 2, 1, 2, \ldots$, and $1, 0, \frac{1}{3}, 0, \frac{1}{3}, \ldots$ and $1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \ldots$ are not. It is easy to see that a sequence is sortable iff every element is strictly smaller than only finitely many elements. It follows that all nonnegative sequences with a finite sum, and hence all probability distributions, are sortable.

**Lemma 10** Let $\overline{a'}$ and $\overline{b'}$ be the sorted versions of two equal-length (possibly infinite) sortable sequences $\overline{a}$ and $\overline{b}$. Then

$$||\overline{a'} - \overline{b'}||_{\infty} \leq ||\overline{a} - \overline{b}||_{\infty}. $$

**Proof** Since $\overline{a}$ and $\overline{b}$ are sortable, $\overline{a'}$ and $\overline{b'}$ are well defined. Furthermore, as mentioned above, it suffices to prove the lemma for sequences $\overline{a}$ that are already sorted, namely $\overline{a'} = \overline{a}$.

Let $S = ||\overline{a} - \overline{b}||_{\infty} \stackrel{\text{def}}{=} \sup_i |a_i - b_i|$. We show that for all $\epsilon > 0$, $\sup_i |a_i - b_i| \geq S - \epsilon$, hence $||\overline{a} - \overline{b}||_{\infty} \geq S$.

Given $\epsilon > 0$, there must be an $i$ such that $|a_i - b_i| \geq S - \epsilon$. There are two possibilities. In the first, $b_i < a_i$. Since $b'$ is sorted, at most $i - 1$ elements of $\overline{b}$ are $> b'_i$. Hence there is a $1 \leq j \leq i$ such that $b_j \leq b'_i$, and since $\overline{a}$ is sorted,

$$\sup_i |a_i - b_i| \geq a_j - b_j \geq a_i - b'_i \geq S - \epsilon.$$

Similarly for $b'_i \geq a_i$. Again since $b'$ is sorted, at least $i$ elements of $\overline{b}$ are $\geq b'_i$. Hence there is a $j \geq i$ such that $b_j \geq b'_i$, and

$$\sup_i |a_i - b_i| \geq b_j - a_j \geq b'_i - a_i \geq S - \epsilon. \quad \square$$
Theorem 11  For every \( f(n) = o(\sqrt{n}) \), SML is uniformly \( f(n) \)-consistent under \( \ell_\infty \) over \( \mathcal{P}_M \).

**Proof**  Since \( f(n) = o(\sqrt{n}) \), for large \( n \), \( 1/f(n) > 4/\sqrt{n} \), and by Lemmas 9 and 10,

\[
P\left(\|N' - P\|_\infty \geq \frac{1}{f(n)}\right) \leq P\left(\|N - P\|_\infty \geq \frac{1}{f(n)}\right) \leq 3 \exp\left(-\frac{n}{8f^2(n)}\right) \to 0. \quad \square
\]

We use this result to show that PML too is a uniformly consistent estimator of \( P \), albeit at a possibly slower rate. Essentially, we have seen that SML will result in an estimate with small \( \ell_\infty \) distance from the underlying \( P \). Hence any given distribution that is far away from \( P \) will assign low probability to the observed pattern. Using the union bound we then multiply by the number of possible PML distributions to see that with high probability, none of those far from \( P \) will assign the pattern a higher probability.

If \( P \) and \( Q \) are two probability distributions over the same domain and \( X \) is distributed according to \( P \), then \( P(Q(X) > P(X)) \) may be arbitrarily high. For example, if \( P \) is \( B(1 - \epsilon) \) and \( Q \) is \( B(1) \), then

\[
P(Q(X) > P(X)) = P(X = 1) = 1 - \epsilon.
\]

However, if \( P \) and \( Q \) have most of their probability on disjoint sets, this cannot happen.

**Lemma 12**  If \( P \) and \( Q \) are probability distributions over a support set \( \mathcal{X} \) and \( X \sim P \), then

\[
P(Q(X) \geq P(X)) \leq \min_{A \subseteq \mathcal{X}} \left( P(A) + Q(A^c) \right).
\]

**Proof**  For every set \( A \),

\[
P(Q(X) \geq P(X)) \leq P(X \in A, Q(X) \geq P(X)) + P(X \in A^c, Q(X) \geq P(X)) \leq P(X \in A) + Q(X \in A^c). \quad \square
\]

**Theorem 13**  For all \( f(n) = o(n^{1/4}) \), PML is uniformly \( f(n) \)-consistent under \( \ell_\infty \) over \( \mathcal{P}_M \).

**Proof**  For \( r \geq 0 \), let

\[
T_P(r) \overset{\text{def}}{=} \{ Q \in \mathcal{P}_M : ||Q - P||_\infty \leq r \},
\]

be the set of distributions whose elements are all within \( r \) from the corresponding elements of \( P \).

Identify a pattern with its sorted SML distribution, for example \( 1232 \) with \( (1, r, 2, r, 3, r, 2, r, 1) \).

Since \( f(n) = o(n^{1/4}) \), for sufficiently large \( n \), \( f(n) \leq \sqrt{n}/8 \), and by Theorem 11, for every \( P \in \mathcal{P}_M \),

\[
P\left(\psi(\overline{X}) \not\in T_P\left(\frac{1}{2f(n)}\right)\right) \leq 3 \exp\left(-\frac{n}{32f^2(n)}\right).
\]

If \( ||Q - P||_\infty > r \), the sets \( T_P(r/2) \) and \( T_Q(r/2) \) are disjoint. Hence if \( ||P - Q||_\infty > \frac{1}{f(n)} \), then \( P \) and \( Q \) assign high probability to distinct patterns, and by Lemma 12, if \( \overline{X} \overset{\text{def}}{=} X_1, \ldots, X_n \sim P \), then

\[
P(Q(\psi(\overline{X})) > P(\psi(\overline{X}))) \leq P\left(T_P^c\left(\frac{1}{2f(n)}\right)\right) + Q\left(T_P\left(\frac{1}{2f(n)}\right)\right) \leq P\left(T_P^c\left(\frac{1}{2f(n)}\right)\right) + Q\left(T_Q\left(\frac{1}{2f(n)}\right)\right) \leq 6 \exp\left(-\frac{n}{32f^2(n)}\right).
\]

Let

\[
\mathcal{P}_n \overset{\text{def}}{=} \{ \hat{P}(\overline{\psi}) : \overline{\psi} \in \mathcal{V}_n \}
\]
be the collection of all maximum-likelihood distributions of length-\(n\) patterns. For example,

\[
P_1 = \{ \hat{P}_1 \} = \mathcal{P}_m, \\
P_2 = \{ \hat{P}_{11}, \hat{P}_{12} \} = \{(1, ()\}, \\
P_3 = \{ \hat{P}_{111}, \hat{P}_{112}, \ldots, \hat{P}_{123} \} = \{(1, (5, 5), ()\}.
\]

The number of distributions in \(P_n\) is at most the number of length-\(n\) profiles, namely the partition number of \(n\). Hence

\[
|P_n| \leq \bar{\psi}_n \leq e^{\sqrt{n}}.
\]

\[\text{write precise}\]

\[
P\left(\|\hat{P}(\mathbf{X}) - P\|_\infty \geq \frac{1}{f(n)}\right) \leq P\left( \exists Q \in \mathcal{P}_n \cap T_P^{(1/f(n))} : Q(\psi(\mathbf{X})) \geq P(\psi(\mathbf{X})) \right)
\]

\[
\leq |P_n| \cdot \sup_{Q \in T_P^{(1/f(n))}} P(\psi(\mathbf{X}) \geq P(\psi(\mathbf{X})))
\]

\[
\leq 6 \exp(\sqrt{n}) \cdot \exp \left( -\frac{n}{32 f^2(n)} \right)
\]

\[
\rightarrow 0
\]

uniformly for all distributions in \(P_m\), as long as \(f(n) = o(n^{1/4})\).

\[\square\]

\section{\(\ell_1\) consistency}

We show that under the \(\ell_1\) norm, the PML is consistent over \(\mathcal{P}_m\). However, unlike in the case of the \(\ell_\infty\) norm, PML is not uniformly consistent.

\textbf{Theorem 14} \hspace{1cm} PML is consistent under \(\ell_1\) over \(\mathcal{P}_m\).

\textbf{Proof} \hspace{1cm} Let \(\mathbf{X} = X_1, \ldots, X_n, X_i \sim P = (p_1, p_2, \ldots),\) be the observed sequence. For any \(\epsilon > 0\), let \(t\) be such that

\[
\sum_{i=t+1}^{\infty} p_i \leq \frac{\epsilon}{4}.
\]

For \(r \geq 0\) and any \(P\), let

\[
T_P(r) \overset{\text{def}}{=} \left\{ Q \in \mathcal{P}_m : \|Q - P\|_\infty \leq r, \left| \sum_{i=t+1}^{\infty} q_i - \sum_{i=t+1}^{\infty} p_i \right| \leq r \right\},
\]

be the set of distributions whose elements are all within \(r\) from the corresponding elements of \(P\), and further their sum is within \(r\) of \(P\)‘s. Observe that for any two distributions \(P\) and \(Q\) and \(t > 0\)

\[
\|P - Q\|_1 = \sum_{i=1}^{\infty} |p_i - q_i|
\]

\[
\leq \sum_{i=1}^{t} |p_i - q_i| + \sum_{i=t+1}^{\infty} p_i + \sum_{i=t+1}^{\infty} q_i
\]

\[
= \sum_{i=1}^{t} |p_i - q_i| + \sum_{i=t+1}^{\infty} p_i + \sum_{i=1}^{\infty} q_i - \sum_{i=1}^{t} q_i
\]

\[
\rightarrow 0
\]
\[
\sum_{i=1}^{t} |p_i - q_i| + \sum_{i=t+1}^{\infty} p_i + \sum_{i=t+1}^{\infty} q_i + \sum_{i=1}^{\infty} p_i - \sum_{i=1}^{\infty} q_i - \sum_{i=1}^{t} (p_i - |p_i - q_i|)
\]

\[
= 2 \sum_{i=1}^{t} |p_i - q_i| + 2 \sum_{i=t+1}^{\infty} p_i + \sum_{i=1}^{\infty} p_i - \sum_{i=1}^{\infty} q_i
\]

\[
\leq 2t\|P - Q\|_\infty + 2 \sum_{i=t+1}^{\infty} p_i + \sum_{i=1}^{\infty} p_i - \sum_{i=1}^{\infty} q_i.
\]

Therefore if \(\|P - Q\|_\infty \leq \frac{\epsilon}{4t+2}\), and

\[
\left| \sum_{i} q_i - \sum_{i} p_i \right| \leq \frac{\epsilon}{4t+2},
\]

then \(\|P - Q\|_1 \leq \epsilon\). Therefore if \(\|P - Q\|_1 > \epsilon\), then \(T_P\left(\frac{\epsilon}{8t+4}\right)\) and \(T_Q\left(\frac{\epsilon}{8t+4}\right)\) are disjoint.

From Lemma 9 and the Chernoff bound, for any distribution \(P\)

\[
P\left(T_P\left(\frac{\epsilon}{8t+4}\right)\right) \leq 3 \exp\left(-\frac{ne^2}{8(8t+4)^2}\right) + 2 \exp\left(-\frac{ne^2}{2(8t+4)^2}\right)
\]

\[
\leq 5 \exp\left(-\frac{ne^2}{8(8t+4)^2}\right).
\]

If \(\|P - Q\|_1 > \epsilon\), then as in the proof of Theorem 13, using Lemma 12

\[
P(Q(\psi(X)) > P(\psi(X))) \leq P\left(T_P\left(\frac{\epsilon}{8t+4}\right)\right) + Q\left(T_P\left(\frac{\epsilon}{8t+4}\right)\right)
\]

\[
\leq P\left(T_P\left(\frac{\epsilon}{8t+4}\right)\right) + Q\left(T_Q\left(\frac{\epsilon}{8t+4}\right)\right)
\]

\[
\leq 10 \exp\left(-\frac{ne^2}{8(8t+4)^2}\right).
\]

Recall that \(\mathcal{P}_n\) is the collection of all maximum-likelihood distributions of length-\(n\) patterns, and \(|\mathcal{P}_n| \leq e\sqrt{n}\). Then

\[
P\left(||\hat{P}_{\psi(X)} - P||_1 \geq \epsilon\right) \leq P\left(\exists Q \in \mathcal{P}_n : ||P - Q||_1 > \epsilon, Q(\psi(X)) \geq P(\psi(X))\right)
\]

\[
\leq |\mathcal{P}_n| \cdot 10 \exp\left(-\frac{ne^2}{8(8t+4)^2}\right)
\]

\[
\leq 10 \exp\left(\sqrt{n}\right) \cdot \exp\left(-\frac{ne^2}{8(8t+4)^2}\right)
\]

\[
\to 0
\]

for any fixed \(P \in \mathcal{P}_t\) and \(\epsilon > 0\). \(\Box\)

Note that unlike in the case of the \(\ell_\infty\) norm, the probability of deviation under the \(\ell_1\) norm depends on the distribution under consideration. In our bounds, this manifests in the form of the parameter \(t\) in the exponent. Note that \(t\) can be arbitrarily large for \(P \in \mathcal{P}_t\) and therefore the bound does not uniformly go to zero.

### 10 Nonuniform \(\ell_1\) consistency

We observed that, unlike for the \(\ell_\infty\) norm, the PML distribution is not uniformly consistent for \(\ell_1\). In fact, we show that no multiset estimator is uniformly consistent.
Theorem 15  There exists no sequence of estimators \( \{ f_n(X_1X_2\ldots X_n) \} \) that is uniformly consistent for \( S(P) \) for all \( P \in \mathcal{P} \).

Proof  Let \( \overline{X} = X_1, X_2, \ldots, X_n \) be drawn \( i.i.d. \) according to some \( P \in \mathcal{P} \). Note that for any \( P_1, P_2 \in \mathcal{P}, \| P_1 - P_2 \|_1 \leq 2 \). If \( E_P[\cdot] \) denotes the expectation with respect to the distribution \( P \), then for any \( c > 0 \), the expected \( \ell_1 \) distance between the estimate \( f_n(\overline{X}) \) and the true distribution \( P \) can be bounded as follows

\[
E_P[\| S(P) - f_n(\overline{X}) \|_1] \leq c \cdot P(\| S(P) - f_n(\overline{X}) \|_1 \leq c) + 2 \cdot P(\| S(P) - f_n(\overline{X}) \|_1 \geq c).
\]

Therefore if

\[
E_P[\| S(P) - f_n(\overline{X}) \|_1] \geq 2c
\]

then

\[
P(\| S(P) - f_n(\overline{X}) \|_1 \geq c) \geq \frac{c}{2-c}.
\]

We will show that for any sequence of estimators \( \{ f_n \} \), there exists a sequence \( \{ P^n \} \) of distributions such that for all sufficiently large \( n \)

\[
E_{P^n}(\| S(P^n) - f_n(\overline{X}) \|_1) \geq \frac{1}{2}.
\]  \hspace{1cm} (10)

which will imply that

\[
P^n(\| S(P^n) - f_n(\overline{X}) \|_1 \geq \frac{1}{4}) \geq \frac{1}{4},
\]

in other words, with constant probability, the estimator with \( n \) samples is bounded away from \( P^n \).

Let \( m_1(n) \leq m_2(n) \) be two numbers much larger than \( n^2 \). For a given \( n \), consider the collection of distributions \( Q_n = \{ Q^i_n \} \cup \{ Q^*_n \} \), where \( \{ Q^i_n \} \) is the set of all uniform distributions over some size-\( m_1(n) \) subset of \( \{1, 2, \ldots, m_2(n)\} \), and \( Q^*_n \) is the uniform distribution over \( \{1, 2, \ldots, m_2(n)\} \). Let \( Q \) be a random distribution selected from the collection \( \{ Q^i_n \} \cup \{ Q^*_n \} \) according to the distribution

\[
P_Q(Q) = \begin{cases} 
\frac{1}{2} & Q = Q^*_n \\
\frac{1}{2} |\{ Q^i_n \}|^{-1} & Q = Q^i_n.
\end{cases}
\]

We will show that for any estimator \( f_n \),

\[
E_Q[\| S(Q) - f_n(\overline{X}) \|_1] \geq \frac{1}{2}.
\]  \hspace{1cm} (11)

In other words, the expected distance between the estimate and the estimand when an \( i.i.d. \) sequence \( \overline{X} \) is drawn according to \( Q \) is at least \( \frac{1}{2} \). This proves the Theorem, since (11) implies that there exists at least one distribution \( P^n \) in \( \{ Q^i_n \} \cup \{ Q^*_n \} \) for which (10) holds.
Observe that for all estimators \( f_n \),
\[
E_Q[||S(Q) - f_n(\overline{X})||_1] = \sum_{\overline{\pi}} P_{Q,\overline{X}}(Q = Q^*_n, \overline{\pi})||S(Q^*_n) - f_n(\overline{X})||_1
+ \sum_{\overline{\pi}} \sum_{Q^*_n} P_{Q,\overline{X}}(Q = Q^*_n, \overline{\pi})||S(Q^*_n) - f_n(\overline{X})||_1
= \sum_{\overline{\pi}} P_{Q,\overline{X}}(Q = Q^*_n, \overline{\pi})||S(Q^*_n) - f_n(\overline{X})||_1
+ \sum_{\overline{\pi}} P_{Q,\overline{X}}(Q \in \{Q^*_n\}, \overline{\pi})||S(Q^*_n) - f_n(\overline{X})||_1
\geq \sum_{\overline{\pi}, \overline{\psi}(\overline{\pi}) = 12\ldots n} P_{Q,\overline{X}}(Q = Q^*_n, \overline{\pi})||S(Q^*_n) - f_n(\overline{X})||_1
+ \sum_{\overline{\pi}, \overline{\psi}(\overline{\pi}) = 12\ldots n} P_{Q,\overline{X}}(Q \in \{Q^*_n\}, \overline{\pi})||S(Q^*_n) - f_n(\overline{X})||_1
\tag{12}
\]
where the second equality holds as for all \( i \neq j \), \( S(Q^*_n) = S(Q^*_n) \). By symmetry for all \( \overline{\pi} \) such that \( \overline{\psi}(\overline{\pi}) = 12\ldots n \), \( P_{Q,\overline{X}}(Q \in \{Q^*_n\}, \overline{\pi}) \) is a constant, and similarly for all \( \overline{\pi} \) such that \( \overline{\psi}(\overline{\pi}) = 12\ldots n \), \( P_{Q,\overline{X}}(Q = Q^*_n, \overline{\pi}) \) is a constant. If \( \overline{\Psi}_{12\ldots n} \) is the set of all \( \overline{\pi} \in [m_2(n)]^n \) with pattern 12\ldots n, we have that for all \( \overline{\pi} \in \overline{\Psi}_{12\ldots n} \)
\[
P_{Q,\overline{X}}(Q \in \{Q^*_n\}, \overline{\psi}(\overline{X}) = 12\ldots n) = |\overline{\Psi}_{12\ldots n}| P_{Q,\overline{X}}(Q \in \{Q^*_n\}, \overline{\pi}) \tag{13}
\]
and
\[
P_{Q,\overline{X}}(Q = Q^*_n, \overline{\psi}(\overline{X}) = 12\ldots n) = |\overline{\Psi}_{12\ldots n}| P_{Q,\overline{X}}(Q = Q^*_n, \overline{\pi}). \tag{14}
\]
Observe that for a given \( n \), the probability that the sequence \( \overline{X} = X_1, X_2, \ldots, X_n \) drawn according to \( P_k \), a uniform distribution with support size \( k \), has pattern 12\ldots n is
\[
P_k(\overline{\psi}(\overline{X}) = 12\ldots n) = \binom{k}{n} \frac{n!}{k^n} \geq \left( 1 - \frac{n}{k} \right)^n \geq e^{-\frac{n^2}{k^2}}.
\]
This quantity can be made arbitrarily close to 1 by choosing a sufficiently large \( k \). Therefore for any \( \delta > 0 \) and sufficiently large \( m_1(n) \), \( m_2(n) \), and \( n \)
\[
P_{Q,\overline{X}}(Q \in \{Q^*_n\}, \overline{\psi}(\overline{X}) = 12\ldots n) \geq \frac{1}{2}(1 - \delta) \tag{15}
\]
and
\[
P_{Q,\overline{X}}(Q = Q^*_n, \overline{\psi}(\overline{X}) = 12\ldots n) \geq \frac{1}{2}(1 - \delta). \tag{16}
\]
Combining (13) with (15) and (14) with (16), and substituting in (12), we obtain for any \( \epsilon > 0 \) and sufficiently large \( m_2(n) \)
\[
E_Q[||S(Q) - f_n(\overline{X})||_1] \geq \frac{1}{2} \frac{1 - \delta}{|\overline{\Psi}_{12\ldots n}|} \sum_{\overline{\pi}, \overline{\psi}(\overline{\pi}) = 12\ldots n} (||S(Q^*_n) - f_n(\overline{X})||_1 + ||S(Q^*_n) - f_n(\overline{X})||_1)
\geq \frac{1}{2} \frac{1 - \delta}{|\overline{\Psi}_{12\ldots n}|} \sum_{\overline{\pi}, \overline{\psi}(\overline{\pi}) = 12\ldots n} (||S(Q^*_n) - f_n(\overline{X})||_1 + ||S(Q^*_n) - f_n(\overline{X})||_1)
\geq \frac{1}{2} \frac{(1 - \delta)(2 - \epsilon)}{\overline{\Psi}_{12\ldots n}} \sum_{\overline{\pi}, \overline{\psi}(\overline{\pi}) = 12\ldots n} 1
= (1 - \delta) \left( 1 - \frac{\epsilon}{2} \right)
\]
where the second inequality follows from the triangle inequality for the $\ell_1$ distance and the third inequality holds as $\|S(Q^*_n) - S(Q^*_m)\|_1$ can be made arbitrarily close to 2 by choosing $m_2(n)$ sufficiently large compared to $m_1(n)$.

## 11 PML as Smoothing

In this section, we show that the high-profile distribution estimator is a smoothed version of the empirical-frequency estimator. More precisely, we prove that for all patterns the empirical frequency majorizes the high-profile distribution. To begin, we formally define majorization.

Given two distributions $P = (p_1, p_2, \ldots)$ and $Q = (q_1, q_2, \ldots)$ from the collection $P_M$, $P$ majorizes $Q$, written $P \gg Q$, if for all $i$

$$\sum_{j=1}^i p_j \geq \sum_{j=1}^i q_j.$$

It is easy to see that any distribution majorizes the uniform distribution over the same number of elements and every convex combination of the two. Viewing uniform distributions as the 'smoothest' for any given support size, we therefore view a distribution majorized by another as also 'smoother'.

In Theorem 16 we show that for all patterns $\bar{\psi}$,

$$\hat{N}(\bar{\psi}) \gg \hat{P},$$

namely the empirical frequency majorizes the high-profile distribution.

**Theorem 16** For all non-trivial patterns $\bar{\psi}$,

$$\hat{N}(\bar{\psi}) \gg \hat{P}.$$  

**Proof** Wolog let $\bar{\psi} = 1^\mu_1 2^\mu_2 \ldots m^\mu_m$ where for all $i \in [m-1]$, $\mu_i \geq \mu_{i+1}$. Then $\hat{N}(\bar{\psi}) = (\frac{\mu_1}{n}, \frac{\mu_2}{n}, \ldots, \frac{\mu_m}{n})$. Let $P = (p_1, p_2, \ldots)$. If $i$ is such that

$$\sum_{j=1}^i p_j > \sum_{j=1}^i \frac{\mu_j}{n},$$

then we will show that there exists another distribution that assigns a larger probability to $\bar{\psi}$.

For $0 \leq \alpha \leq 1$, define

$$Q^\alpha = \left(\frac{\alpha}{\sum_{j=1}^i p_j}, \frac{\alpha}{\sum_{j=1}^i p_j}, \ldots, \frac{\alpha}{\sum_{j=1}^i p_j}, \frac{1 - \alpha}{1 - \sum_{j=1}^i p_j}, p_{i+1}, \ldots\right)$$

to be the distribution comprising of two parts: one with the $i$ largest probabilities of $P$ suitably scaled so that they sum to $\alpha$ and the second consisting of the remaining probabilities suitably scaled to sum to $1 - \alpha$. Note that

$$Q^1 = \left(\frac{p_1}{\sum_{j=1}^i p_j}, \frac{p_2}{\sum_{j=1}^i p_j}, \ldots, \frac{p_i}{\sum_{j=1}^i p_j}\right)$$

the distribution consisting of the $i$ largest probability elements in $P$, suitably normalized,

$$Q^0 = \left(\frac{p_{i+1}}{1 - \sum_{j=1}^i p_j}, \frac{p_{i+2}}{1 - \sum_{j=1}^i p_j}, \ldots\right).$$
the distribution consisting of all but the $i$ largest probability elements in $P$, suitably normalized, and that

$$Q^{\sum_{j=1}^i p_j} = P.$$  

For all $\mathcal{X} \subseteq [m]$, let $\bar{\psi}_\mathcal{X}$ denote the pattern obtained by retaining only the members of $\mathcal{X}$ in $\bar{\psi}$. For example if $\mathcal{X} = \{1, 3, 4\}$ then

$$\bar{\psi}_\mathcal{X} = 1^\mu_1 2^\mu_2 3^\mu_4.$$  

For all $S = \{s_1, s_2, \ldots\} \subseteq [k]$ and any distribution $P = (p_1, p_2, \ldots, p_k)$ let $P_S$ denote the distribution

$$P_S = \left(\frac{p_{s_1}}{p_S}, \frac{p_{s_2}}{p_S}, \ldots\right)$$

where $s_1 \geq s_2 \geq \ldots$ and $p_S = \sum_{s \in S} p_s$. This distribution has zero continuous probability. Furthermore let

$$P_{\bar{S}} = \left(\frac{p_{\bar{s}_1}}{1-p_S}, \frac{p_{\bar{s}_2}}{1-p_S}, \ldots\right)$$

where $\bar{S} = \mathbb{P} - S = \{\bar{s}_1, \bar{s}_2, \ldots\}$ and $\bar{s}_i \geq \bar{s}_{i+1}$ for all $i$. Its continuous probability is $q = q_{p_S}$ where $q$ is the continuous probability of $P$. Then for any $\mathcal{X} \subseteq [k]$,

$$P(\bar{\psi}) = \sum_{\mathcal{X} \subseteq [m]} (p_S)^{\sum_{i \in \mathcal{X}} \mu_i} (1-p_S)^{n-\sum_{i \in \mathcal{X}} \mu_i} P_S(\bar{\psi}_\mathcal{X}) P_{\bar{S}}(\bar{\psi}_{\bar{\mathcal{X}}})$$

where $\bar{\mathcal{X}} = [m] - \mathcal{X}$ is the set of integers in $[m]$ but not in $\mathcal{X}$.

Observe that if $S = [i]$ then

$$Q_S^S = Q^1,$$

$$Q_S^\bar{S} = Q^0,$$

and $q_S^S = \alpha$. Hence

$$Q^\alpha(\bar{\psi}) = \sum_{\mathcal{X} \subseteq [m]} \alpha^{\sum_{i \in \mathcal{X}} \mu_i} (1-\alpha)^{n-\sum_{i \in \mathcal{X}} \mu_i} Q^1(\bar{\psi}_\mathcal{X}) Q^0(\bar{\psi}_{\bar{\mathcal{X}}})$$

$$= \sum_{\mathcal{X} \subseteq [m]: |\mathcal{X}| \leq i} \alpha^{\sum_{i \in \mathcal{X}} \mu_i} (1-\alpha)^{n-\sum_{i \in \mathcal{X}} \mu_i} Q^1(\bar{\psi}_\mathcal{X}) Q^0(\bar{\psi}_{\bar{\mathcal{X}}}),$$

where the second equality follows from the fact that $Q^1(\bar{\psi}_\mathcal{X}) = 0$ if $|\mathcal{X}| > i$. The derivative of $\alpha^i(1 - \alpha)^{n-i}$ with respect to $\alpha$ is negative for all $\alpha > i/n$. Therefore the derivative of $Q^\alpha(\bar{\psi})$ with respect to $\alpha$ is negative for all

$$\alpha > \max_{\mathcal{X} \subseteq [m]: |\mathcal{X}| \leq i} \frac{1}{n} \sum_{j \in \mathcal{X}} \mu_j = \frac{1}{n} \sum_{j=1}^i \mu_j.$$  

Since

$$\sum_{j=1}^i p_j > \frac{1}{n} \sum_{j=1}^i \mu_j,$$

$$Q^{\frac{1}{n} \sum_{j=1}^i \mu_j}(\bar{\psi}) > Q^{\sum_{j=1}^i p_j}(\bar{\psi}) = P(\bar{\psi}).$$  

\[\square\]
Acknowledgements

We thank Warwick DeLauney and Doug Jungreis for simplifying our original proof for the high-profile distribution of the pattern 112, Shengjun (Alan) Pan for help with the experiments, and Ian Abramson, Noga Alon, Ron Graham, Nikola Jevtić, and Sajama for helpful discussions.

12 Appendix

12.1 Consistency results for the $\ell_1$ norm

In this Appendix, we derive consistency results for the high-profile distribution when the norm in question is the $\ell_1$ norm. We show that $\hat{P}$ is consistent for $S(P), P \in \mathcal{P}_b$, but is not uniformly consistent. The proof of the first result is similar to that of Theorem 13.

Let $X = X_1X_2\ldots X_n$ be the observed sequence drawn according to $P = (p_1, p_2, \ldots) \in \mathcal{P}_{M,D}$. Let $\tau$ be some integer. Suppose $Q = (q_1, q_2, \ldots) \in \mathcal{P}_{M,D}$ is another distribution and there exists $i \in [\tau]$ such that $|p_i - q_i| > 2\epsilon$ for some $\epsilon > 0$. Then the following is true.

**Lemma 17** For all sufficiently large $n$ there exists a $c > 0$ such that for all $P, Q \in \mathcal{P}_{M,D}$, satisfying the above condition

$$P(Q(\psi(X)) > P(\psi(X))) \leq e^{-cn}.$$  

**Proof** Wolog let $P$'s support be the positive integers $\mathbb{P}$, i.e., for all $i \geq 1$, $P(X_1 = i) = p_i$. Let $N_i = \frac{1}{n} \sum_{j=1}^{n} 1(X_j = i)$ denote the fraction of times $i$ appears in the sequence $X$. Let $\{\mu(i)\}$ denote the multiplicities of a pattern $\hat{\psi}$ in sorted order, i.e., $\mu(i) \geq \mu(i+1)$. The pattern $\hat{\psi}$ is typical for $P \in \mathcal{P}_M$, if for all $i \in [\tau]$,

$$|\frac{\mu(i)}{n} - p_i| < \epsilon.$$  

For any given distribution $P \in \mathcal{P}_M$, the set of all typical $\hat{\psi}$'s is denoted by $T_P$. We will show that if there exists $i \leq \tau$ such that

$$|\frac{\mu(i)}{n} - p_i| \geq \epsilon$$  

then there exists $j$ such that

$$|N_j - p_j| \geq \epsilon.$$  

(17)

(18)

Let $i$ be the smallest integer for which Equation (17) holds and let $\frac{\mu(i)}{n} - p_i \geq \epsilon$. We will show the existence of $j$ satisfying (18). A similar argument applies to the case $p_i - \frac{\mu(i)}{n} \geq \epsilon$. Let $\mu(i)$ correspond to the multiplicity of $j$, i.e., $\frac{\mu(i)}{n} = N_j$. If $j = i$ then $|N_i - p_i| \geq \epsilon$. If $j > i$, then

$$N_j - p_j \geq N_j - p_i = \frac{\mu(i)}{n} - p_i \geq \epsilon.$$  

If $j < i$, note that $i - 1$ symbols appear at least $\mu(i)$ times but at most $i - 2$ of those have probability greater than $p_i$. Hence there exists $t \geq i$ that appears more than $\mu(i)$ times, i.e., $N_t \geq \frac{\mu(i)}{n}$. Therefore

$$N_t - p_t \geq N_t - p_i \geq \frac{\mu(i)}{n} - p_i \geq \epsilon.$$
This implies that for all \( P \in \mathcal{P}_{m,D} \) and all sufficiently large \( n \)
\[
P(\psi(\overline{X}) \notin T_P) = P\left( \exists i \in [\tau] : \left| \frac{\mu(i)}{n} - p_i \right| \geq \epsilon \right) \leq P(\exists j : |N_j - p_j| \geq \epsilon) \leq e^{-cn} \tag{19}
\]
where the last inequality follows from Lemma 9. Note that
\[
P(Q(\psi(\overline{X})) > P(\psi(\overline{X}))) \leq P(\psi(\overline{X}) \in T_P, Q(\psi(\overline{X})) > P(\psi(\overline{X}))) + P(\psi(\overline{X}) \notin T_P) \leq P(\psi(\overline{X}) \in T_P, Q(\psi(\overline{X})) > P(\psi(\overline{X}))) + e^{-cn}
\]
where we used (19)

Therefore to prove the lemma it suffices to show that the first term can be bounded by a term of the form \( e^{-cn} \). Observe that
\[
P(\psi(\overline{X}) \in T_P, Q(\psi(\overline{X})) > P(\psi(\overline{X}))) < Q(\psi(\overline{X}) : \psi(\overline{X}) \in T_P),
\]
we will bound the latter.

Let \( j \) be such that \( |q_j - p_j| > 2\epsilon \). Then by the definition of typical patterns
\[
|T_P \cap T_Q| = 0.
\]
Therefore
\[
P(\psi(\overline{X}) \in T_P, Q(\psi(\overline{X})) > P(\psi(\overline{X}))) < Q(\psi(\overline{X}) \in T_P) \leq Q(\psi(\overline{X}) \notin T_Q) \leq e^{-cn}
\]
for some \( c > 0 \), where the last inequality follows from the application of (19) to the distribution \( Q \).

Now we are in a position to prove the consistency of high-profile distributions for distributions in \( \mathcal{P}_D \).

**Theorem 18** The sequence \( \{\hat{P}_n\} \) of estimators is consistent for \( P \in \mathcal{P}_D \).

**Proof** Let \( \overline{X} = X_1 X_2 \ldots X_n \) be the observed sequence. As \( \hat{P}_n \) is a function of the profile of \( \overline{X} \), for any two distributions \( P_1, P_2 \in \mathcal{P}_D \) such that \( S(P_1) = S(P_2) \), the distributions of \( ||\hat{P}_n - S(P_1)|| \) and \( ||\hat{P}_n - S(P_2)|| \) are identical. Therefore it suffices to prove consistency result for \( P \in \mathcal{P}_{m,D} \).

More precisely, we need to show that for all \( P \in \mathcal{P}_{m,D} \),
\[
\hat{P}_n \xrightarrow{P} P.
\]

Let \( P = (p_1, p_2, \ldots) \) be the true distribution. Since \( \sum_i p_i = 1 \) and \( p_i \geq p_{i+1} \), given \( \epsilon > 0 \), one can choose \( \tau \) such that
\[
\sum_{i=\tau+1}^{\infty} p_i \leq \frac{\epsilon}{4} . \tag{20}
\]

Let
\[
\mathcal{P}_P \overset{\text{def}}{=} \left\{ Q = (q_1, q_2, \ldots) \in \mathcal{P}_m : \exists i \in [\tau] \text{ such that } |q_i - p_i| \geq \frac{\epsilon}{4\tau} \right\}
\]
be the collection of distributions where at least one of the \( q_i, 1 \leq i \leq \tau \), differs from \( p_i \) by more than \( \frac{\epsilon}{4\tau} \).

The set \( \mathcal{P}_P \) is compact as it is a closed subset of the compact set \( \mathcal{P}_m \). Therefore for every pattern \( \tilde{\psi} \) there exists \( \hat{P}_{\tilde{\psi}} \in \mathcal{P}_P \) such that
\[
\hat{P}_{\tilde{\psi}}(\psi) = \sup_{Q \in \mathcal{P}_P} Q(\psi).
\]
Let 
\[ \tilde{\mathcal{P}} \overset{\text{def}}{=} \{ \tilde{\mathcal{P}} \} \]
denote the collection of all distributions that achieve the supremum for some \( \tilde{\psi} \), where if for some pattern the maximizing distribution is not unique, we select one of them. Recall that all patterns of the same profile are assigned the same probability, and therefore we can select one distribution for all the patterns with the same profile. Hence the number \( |\tilde{\mathcal{P}}| \) of such distributions selected is at most the number of profiles of length \( n \). It was shown in [?] that this number is \( e^{O(\sqrt{n})} \).

Recall that \( \hat{\mathcal{P}}_n = (\hat{p}_1, \hat{p}_2, \ldots) \) is the high-profile distribution of the pattern \( \tilde{\psi} \). Let 
\[ S_1 \overset{\text{def}}{=} \{ \tilde{\psi} : \exists i \in [\tau] \text{ such that } |\hat{p}_i - p_i| \geq \frac{\epsilon}{4\tau} \} \]
denote the set of patterns whose \( \hat{p}_i \) differs markedly from \( p_i \) for at least one \( i \in [\tau] \). Then \( \tilde{\psi} \in S_1 \) implies that \( P(\tilde{\psi}) < \hat{P}(\tilde{\psi}) \) and hence

\[
P(\tilde{\psi} \in S_1) \leq P\left( \hat{P}(\psi(X)) > P(\psi(X)) \right) \leq \sum_{Q \in \mathcal{P}} P(Q(\psi(X)) > P(\psi(X))).\]

From Lemma 17 for every \( Q \in \mathcal{P}_P \)
\[
P(Q(\psi(X)) > P(\psi(X))) \leq e^{-cn},
\]
where \( c > 0 \) is a function of \( \tau \), and \( \epsilon \). Hence

\[
P(\tilde{\psi} \in S_1) \leq \sum_{Q \in \mathcal{P}} P(Q(\psi(X)) > P(\psi(X))) \leq \sum_{Q \in \mathcal{P}} e^{-cn} \leq e^{-(cn - c_1 \sqrt{n})},\]

where \( e^{c_1 \sqrt{n}} \) is an upper bound on the number of profiles of length \( n \). Therefore for any \( \delta \), by choosing a sufficiently large \( n \) we obtain \( P(\tilde{\psi} \in S_1) \leq \delta \). This implies that with probability at least \( 1 - \delta \)
\[
\sum_{i=1}^{\tau} |p_i - \hat{p}_i| \leq \sum_{i=1}^{\tau} \frac{\epsilon}{4\tau} = \frac{\epsilon}{4}. \quad (21)
\]

Also,
\[
\sum_{i=\tau+1}^{\infty} \hat{p}_i \leq 1 - \sum_{i=1}^{\tau} \hat{p}_i \\
\leq \sum_{i=\tau+1}^{\infty} p_i + \sum_{i=1}^{\tau} p_i - \sum_{i=1}^{\tau} \hat{p}_i \\
\leq \frac{\epsilon}{4} + \sum_{i=1}^{\tau} |p_i - \hat{p}_i| \\
\leq \frac{\epsilon}{2}
\]
where we used (21) in the last inequality. Therefore for any $\delta > 0$, for sufficiently large $n$, with probability at least $1 - \delta$,

$$||\hat{P}_n - P||_1 \leq \sum_{i=1}^{\tau} |p_i - \hat{p}_i| + \sum_{i=\tau+1}^{\infty} p_i + \sum_{i=\tau+1}^{\infty} \hat{p}_i \leq \epsilon,$$

i.e.,

$$\hat{P} \xrightarrow{P} P.$$

13 Introduction

In this paper we determine the pattern maximum-likelihood (PML) estimate (also referred to as the high-profile distribution) for some simple and short data samples, and establish some of its properties for more complex samples.

The properties of high-profile distributions can be broadly divided into two categories. In Section 14, we consider several properties of high-profile distributions: their total support size, discrete size, continuous probability, and an ordering of patterns based on their high-profile probabilities. These results do not assume any specific structure to the pattern in question. Deriving the high-profile distribution of profiles in general, appears to be a hard problem, and therefore in Section 15 we consider some simple profiles, whose structure makes the problem somewhat tractable. We describe the results in each of these sections in greater detail below.

The existence and probability of the continuous part of the high-profile distribution are considered in Section 14.1. Clearly, if $\varphi_1 = 0$, namely every symbol appears at least twice, then $\hat{P}$ is discrete, namely has no continuous part. We show that this holds also when $\varphi_1 = 1$, namely, exactly one of the symbols appears once. A more refined question concerns the continuous probability $\hat{q}$. In Theorem 19, we provide a simple bound on the continuous probability, showing that for all nontrivial profiles,

$$\hat{q} \leq \frac{\varphi_1}{n},$$

namely, the continuous probability of the high-profile distribution is at most the fraction of elements appearing once.

The example of ten elements, each appearing twice, showed that the discrete support of the high-profile distribution can be larger than the number of symbols observed in the data. In subsections 14.2 to 14.5 we evaluate its possible sizes.

We begin in subsection 14.2 by showing that the discrete size of the high-profile distribution is always finite. This result is also useful as it allows us later to assume a concrete (finite) form of the high-profile distribution.

In subsection 14.4 we refine the finitude result by upper bounding the size of the high-profile support. Recall that $m$ is the number of symbols, and that $\mu_{\min}$ is the smallest number of times any symbol appears in a pattern. In Theorem 26 we show that for all nontrivial profiles,

$$\hat{t} \leq m + \frac{m - 1}{2^\mu_{\min} - 2},$$

implying also the corresponding bound on $\hat{k}$. In particular, if

$$\mu_{\min} > \log(m + 1),$$

36
Table 2: High-profile distribution of profiles of length at most 4.

<table>
<thead>
<tr>
<th>$\bar{\varphi}$</th>
<th>Canonical $\bar{\psi}$</th>
<th>$\hat{P}_{\bar{\varphi}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1^1$</td>
<td>1</td>
<td>any distribution</td>
</tr>
<tr>
<td>$2^1, 3^1, 4^1$</td>
<td>11, 111, 1111</td>
<td>(1)</td>
</tr>
<tr>
<td>$1^2, 1^3, 1^4$</td>
<td>12, 123, 1234</td>
<td>()</td>
</tr>
<tr>
<td>$2^{11^1}$</td>
<td>112</td>
<td>(1/2, 1/2)</td>
</tr>
<tr>
<td>$3^11^1$</td>
<td>1112</td>
<td>(1/2, 1/2)</td>
</tr>
<tr>
<td>$2^2$</td>
<td>1122</td>
<td>(1/2, 1/2)</td>
</tr>
<tr>
<td>$2^{11^2}$</td>
<td>1123</td>
<td>(1/5, 1/5, 1/5, 1/5, 1/5)</td>
</tr>
</tbody>
</table>

\[ \hat{k} = \hat{t} = m, \]  

namely, if the number of times each symbol appears is at least the logarithm of the number of symbols, then the high-profile support does not use any additional symbols.

In subsection 14.5 we turn to lower bounds on $\hat{t}$. Recall that $\mu_{\text{max}}$ is the maximum number of times any symbol appears. In Theorem 30 we show that

\[ \hat{t} \geq m - 1 + \left( \frac{\sum \hat{\varphi}_\mu \cdot 2^{-\mu} - 2^{-\mu_{\text{max}}}}{2^{\mu_{\text{max}}}} \right). \]  

(24)

In particular, complementing (23), if

\[ \mu_{\text{max}} < \log \left( \sqrt{m} + 1 \right), \]

namely, the number of times each symbol appears is at most half the logarithm of the number of symbols, then

\[ \hat{t} > m, \]

namely, the high-profile support is larger than the number of symbols observed.

In Subsections 15.1 to 15.4, we use and extend the results hitherto obtained to determine the high-profile distribution of some short and simple profiles.

In Subsection 15.1, we determine $\hat{P}$ for all profiles of length at most 4. For example, we show that the high-profile distribution of the profile $2^{11^1}$, corresponding to the sequence @@@# mentioned in the introduction, is indeed $(1/2, 1/2)$. Perhaps the most “interesting” of these profiles is $2^{11^2}$, where one symbol appears twice and two symbols appear once, as in the sequence @#$@$. This is the shortest profile for which the high-profile discrete support size—5, is larger than the number of symbols appearing—3. Table 2 summarizes the results proven in this subsection.

In Subsection 15.2, we consider profiles of the form $\bar{\varphi} = r^1u^u$, namely, one symbol appears $r$ times and $u$ symbols appear once each. We show that for certain values of $r, u$, $\hat{P}_{\bar{\varphi}} = \left( \frac{r}{r+u} \right)$, a mixed distribution comprising a single symbol with probability $\frac{r}{r+u}$ and a continuous part with probability $\frac{u}{r+u}$.

In Subsection 15.3, we consider the high-profile distribution of the quasi-uniform profile where each of $m$ symbols repeats $r$ or $r+1$ times, i.e., profiles of the form $\bar{\varphi} = (r+1)^{m-t}r^t$.  

37
Figure 6: High-profile distribution of binary sequences for increasing data lengths. X axis: Fraction of 0s. Y axis: \( p \) where the High-Profile distribution is \((p, 1 - p)\). Dotted line: \( n = 10 \), dashed line: \( n = 100 \), solid line \( n = \infty \).

We show that \( \hat{P} \) is uniform and distributed over
\[
\hat{k} = \min \left\{ k \geq m : \left( 1 + \frac{1}{k} \right)^{rm+t} \frac{m}{k+1} > 1 \right\}.
\]
symbols. In particular, as mentioned in the introduction, \( \hat{k} = 12 \), when \( m = 10 \), \( r = 2 \), and \( t = 0 \). We show that follows that for fixed \( r \), as \( m \) grows, \( \hat{k}/m \) approaches the solution of
\[
x \log \frac{x}{x - 1} = r.
\]
For example, for \( r = 2 \), the ratio \( \hat{k}/m \) approaches 1.258..., hence the discrete support is a constant factor larger than the number of symbols observed. We also show that if \( r = 1 \) and \( t \geq 1 \) is a constant, then
\[
\lim_{m \to \infty} \frac{t_{(r+1)t(m), r-m-r(t(m))}}{m^2} = \frac{1}{2t}.
\]

In subsection 15.4 we consider sequences consisting of just two symbols. Let \( n_0 \) and \( n_1 \), where wolog \( 1 \leq n_0 \leq n_1 \leq n - 1 \), denote the number of times each of the two symbols appears in a sequence of length \( n \). Combining the upper bound (22) with a result by Alon [?], we show in Theorem 49 that if \( (n_1 - n_0)^2 \leq n \), then \( \hat{P} = (0.5, 0.5) \), and if \( (n_1 - n_0)^2 > n \), then \( \hat{P} = (\frac{1}{1+\alpha}, \frac{\alpha}{1+\alpha}) \), where \( \alpha \) is the unique root in \((0,1)\) of the polynomial
\[
n_0 \cdot x^{n_1-n_0+1} - n_1 \cdot x^{n_1-n_0} + n_1 \cdot x - n_0.
\]

Figure 13 plots the probability \( p = \alpha/(\alpha + 1) \) indicated by the high-profile distribution as a function of \( n_0/n \) for various sequence lengths \( n \). Observe that \( p \) is always between \( n_0/n \) and \( \frac{1}{2} \).

As Figure 13 shows, while the high-profile and maximum-likelihood distributions of binary sequences differ for every block length \( n \), as \( n \) tends to infinity, they converge to each other.

14 Some general properties of high-profile distributions

We prove several results on several quantities related to high-profile distributions: their total support size, discrete size, continuous probability, and an ordering of patterns based on their high-profile probabilities. These results do not assume any specific structure to the pattern in question.

14.1 Continuous probability

One of the first steps in deriving a pattern’s MLD \( \hat{P} \) is finding its continuous probability \( \hat{q} \), which in particular determines whether \( \hat{P} \) is discrete, continuous, or mixed. We first upper \( \hat{q} \) in terms of \( \varphi_1 \), the number of symbols appearing once, showing that for all nontrivial (i.e., length at least two) patterns,
\[
\hat{q} \leq \frac{\varphi_1}{n}.
\]
This clearly implies that all patterns with \( \varphi_1 = 0 \), namely all symbols appear at least twice, have discrete high-profile distributions. We then show that the same holds for all patterns with \( \varphi_1 = 1 \), namely exactly one symbol appears once.

**Theorem 19**  For all non-trivial patterns,

\[
\hat{q} \leq \frac{\varphi_1}{n}.
\]

**Proof** Let \( \tilde{\psi} \) be a non-trivial pattern where \( \varphi_1 \) symbols appear once and let \( P = (p_1, p_2, \ldots) \) have a continuous probability

\[
q \overset{\text{def}}{=} 1 - \sum_i p_i > \frac{\varphi_1}{n}.
\]

We show that another distribution assigns to \( \tilde{\psi} \) a higher probability.

For \( 0 \leq \alpha \leq 1 \), let \( Q^\alpha = \left( \frac{1 - \alpha}{1 - q} p_1, \frac{1 - \alpha}{1 - q} p_2, \ldots \right) \) be the distribution consisting of a continuous probability \( \alpha \), and the suitably normalized discrete probabilities of \( P \). Note that \( Q^q = P \), and that \( Q^0 = \left( \frac{p_1}{1 - q}, \frac{p_2}{1 - q}, \ldots \right) \) is a discrete distribution consisting of the normalized discrete probabilities of \( P \).

Recall that without loss of generality we consider only canonical patterns, which are therefore of the form

\[
\tilde{\psi} = \psi_1 \ldots \psi_n = 1^{\mu_1} 2^{\mu_2} \ldots (m - \varphi_1)^{\mu_m-\varphi_1} (m - \varphi_1 + 1) \ldots m,
\]

where the last \( \varphi_1 \) symbols have multiplicity 1. For \( 0 \leq i \leq \varphi_1 \) let

\[
\psi_i^{n-i} \overset{\text{def}}{=} \psi_1 \ldots \psi_{n-i}
\]
denote the pattern obtained by omitting the last \( i \) symbols of \( \tilde{\psi} \). For example if \( \tilde{\psi} = 11122345 \) then \( \psi^5 = \tilde{\psi}, \psi^7 = 1112234, \) and \( \psi^8 = 111223 \).

Let \( i \) denote the number of times the continuous part appears in the observed sequence. For \( 0 \leq i \leq \varphi_1 \),

\[
Q^\alpha (\tilde{\psi}, i) = \binom{\varphi_1}{i} \alpha^i (1 - \alpha)^{n-i} Q^0 (\psi^{n-i}).
\]

For any fixed \( i \), \( \alpha^i (1 - \alpha)^{n-i} \) decreases with \( \alpha > i/n \), and hence so does

\[
Q^\alpha (\tilde{\psi}) = \sum_{i=0}^{\varphi_1} Q^\alpha (\tilde{\psi}, i).
\]

Since we assumed that \( q > \varphi_1/n \),

\[
P(\tilde{\psi}) = Q^q (\tilde{\psi}) < Q^{\varphi_1/n} (\tilde{\psi}).
\]

The next corollary follows.

**Corollary 20**  All non-trivial patterns with \( \varphi_1 = 0 \) have discrete MLD.
When $\varphi_1 = 1$, namely exactly one symbol appears once, the theorem implies only that $\hat{q} \leq 1/n$. However, we now show that here too the MLD is discrete, hence $\hat{q} = 0$.

**Theorem 21** All non-trivial patterns with $\varphi_1 = 1$ have discrete MLD.

**Proof** Let $\tilde{\psi}$ have $\varphi_1 = 1$. We show that for any distribution $P = (p_1, p_2, \ldots)$ with continuous probability $q > 0$, there is another distribution $Q$, assigning to $\tilde{\psi}$ a higher probability.

If $P$ has $k < m - 1$, then $P(\tilde{\psi}) = 0$, and we can take $Q$ to be any distribution on $m$ or more discrete symbols. We therefore assume that $k \geq m - 1$.

Let $Q = (p_1, p_2, \ldots, q)$ be the distribution obtained by replacing the continuous part of $P$ with a single discrete value of equal probability. Slightly abusively let $x^{n-1} \neq q$ indicate that none of $x_1, \ldots, x_{n-1}$ is the new discrete symbol of probability $q$, and define

$$Q(\tilde{\psi}, X^{n-1} \neq q) \overset{\text{def}}{=} Q(\{x = x_1 x_2 \ldots x_n : \psi(x) = \tilde{\psi}, x^{n-1} \neq q\})$$

to be the probability that $Q$ generates $n$ elements with pattern $\tilde{\psi}$ such that none of the first $n - 1$ is the new discrete symbol. Similarly let

$$Q_{-q}(\tilde{\psi}) \overset{\text{def}}{=} Q(\tilde{\psi}, X \neq q),$$

and recalling that $X$ denotes the discrete part of $P$,

$$P_{-q}(\tilde{\psi}) \overset{\text{def}}{=} P(\{x = x_1 x_2 \ldots x_n : \psi(x) = \tilde{\psi}, x_1, x_2, \ldots, x_n \notin X\}).$$

Note that $Q_{-q}(\tilde{\psi}) = P_{-q}(\tilde{\psi})$ for every pattern $\tilde{\psi}$. Assuming as usual that $\tilde{\psi}$ is canonical,

$$Q(\tilde{\psi}) > Q(\tilde{\psi}, X^{n-1} \neq q)
= Q_{-q}(\tilde{\psi}) + Q_{-q}(\tilde{\psi}^{n-1}) \cdot q
= P_{-q}(\tilde{\psi}) + P_{-q}(\tilde{\psi}^{n-1}) \cdot q
= P(\tilde{\psi}),$$

where the inequality follows from $k \geq m - 1$. \hfill $\square$

### 14.2 Finite discrete size

We prove that the discrete size $\hat{k}$ of the high-profile distribution is always finite. First, we provide a simple proof of this result for all patterns with $\varphi_1 = 0$. We then show that the number $|\{p_1, p_2, \ldots\}|$ of distinct probabilities in the high-profile distribution of a pattern is at most its length. It follows that the high-profile distribution has finite discrete size.

**Should this Theorem be removed?**

**Theorem 22** All non-trivial patterns with $\varphi_1 = 0$ have finite high-profile support.

**Proof** From Corollary 20, $\hat{P}$ is discrete. To show that it has finite support, we prove that any distribution with infinite support can be replaced by a distribution with finite support that assigns a greater probability to the pattern.

Let the pattern $\tilde{\psi} = \psi_1 \psi_2 \ldots \psi_n$ have $\varphi_1 = 0$, namely all symbols appear at least twice. Let $P = (p_1, p_2, \ldots) \in \mathcal{P}$ be any discrete distribution with infinite support. Choose a sufficiently large $t$ such that

$$\alpha \overset{\text{def}}{=} \sum_{\tau=t}^{\infty} p_{\tau}$$
satisfies
\[ 1 - \frac{m \alpha^2}{P(\bar{\psi})} > (1 - \alpha)^n. \tag{25} \]

Since for small \( \alpha \), \((1 - \alpha)^n \) is approximately 1 – \( n \alpha \), one can always find such a \( t \). Let \( \bar{X} = X_1 X_2 \ldots X_n \) be the sequence of observed symbols. Let \( E \) be the event that
\[ \exists i \in [n], X_i \in \{t, t + 1, \ldots\}, \]

namely some element of \( \bar{X} \) arises from the tail of the distribution. Then
\[ P(\bar{\psi}) = P(\psi(\bar{X}) = \bar{\psi}) = P(\psi(\bar{X}) = \bar{\psi}, E) + P(\psi(\bar{X}) = \bar{\psi}, \bar{E}). \tag{26} \]

For all \( i \in [m] \), let
\[ L_i \overset{\text{def}}{=} \{ j : \psi_j = i \} \]

be the set of locations where \( i \) occurs in \( \bar{\psi} \). Then the first term in (26) can be bounded by
\[ P(\psi(\bar{X}) = \bar{\psi}, E) = P(\exists i \in [m], \tau \in \{t, t + 1, \ldots\} : \forall j \in L_i, X_j = \tau \text{ and } \psi(\bar{X}) = \bar{\psi}) \]
\[ \leq P(\exists i \in [m], \tau \in \{t, t + 1, \ldots\} : \forall j \in L_i, X_j = \tau) \]
\[ \leq \sum_{i=1}^{m} \sum_{\tau=t}^{\infty} P(\forall j \in L_i, X_j = \tau) \]
\[ = \sum_{i=1}^{m} \sum_{\tau=t}^{\infty} p_i^{\mu_i} \]
\[ \leq \sum_{i=1}^{m} \sum_{\tau=t}^{\infty} p_i^{2} \]
\[ \leq m \left( \sum_{\tau=t}^{\infty} p_\tau \right)^2 \]
\[ \leq m \alpha^2. \tag{27} \]

To bound the second term in (26), construct a new distribution \( Q = (q_1, q_2, \ldots, q_{t-1}) \) by truncating the tail of \( P \) and normalizing the first \( t - 1 \) probabilities to sum to one, namely, for \( i \in [t - 1], \)
\[ q_i = \frac{p_i}{1 - \alpha}. \]

Since for all \( i \in [t - 1], p_i = (1 - \alpha) q_i, \)
\[ P(\psi(\bar{X}) = \bar{\psi}, \bar{E}) = P(\forall j \in [n], X_j \in [t - 1] \text{ and } \psi(\bar{X}) = \bar{\psi}) \]
\[ = \sum_{\pi \in [t - 1]^n : \psi(\pi) = \bar{\psi}} P(\pi) \]
\[ = (1 - \alpha)^n \sum_{\pi \in [t - 1]^n : \psi(\pi) = \bar{\psi}} Q(\pi) \]
\[ = (1 - \alpha)^n Q(\bar{\psi}). \tag{28} \]

Substituting (27) and (28) into (26) we obtain
\[ P(\bar{\psi}) \leq m \alpha^2 + (1 - \alpha)^n Q(\bar{\psi}). \]

And, using (25),
\[ Q(\bar{\psi}) \geq \frac{1 - \frac{m \alpha^2}{P(\bar{\psi})}}{(1 - \alpha)^n} P(\bar{\psi}) > P(\bar{\psi}). \]
To extend this result to all patterns, we first upper bound the number of distinct probability values in any high-profile distribution. Let

$$V(P) \overset{\text{def}}{=} |\{p_1, p_2, \ldots\}|,$$

denote the number of distinct probabilities in $P = (p_1, p_2, \ldots)$.

**Theorem 23** For any pattern $\bar{\psi}$ with high profile distribution $\hat{P}$,

$$V(\hat{P}) \leq \min\{2^m, n-1\}.$$

**Proof** We show that all possible probability values of a high-profile distribution are roots of a low-degree polynomial and use that fact to upper bound their number. Without loss of generality assume that $\bar{\psi} = 1^{\mu_1}2^{\mu_2} \ldots m^{\mu_m}$ is canonical. Let $\hat{\bar{\psi}} = (\hat{p}_1, \hat{p}_2, \ldots)$ be a high-profile distribution of $\bar{\psi}$.

From Kuhn-Tucker conditions for some $\lambda$,

$$\frac{\partial P(\bar{\psi})}{\partial p_j} \bigg|_{P=\hat{P}} = \lambda$$

for all $j \geq 1$.

For $i \in [m]$ let

$$\bar{\psi}_i \overset{\text{def}}{=} 1^{\mu_1} \ldots (i-1)^{\mu_i-1} i^{\mu_{i+1}} \ldots (m-1)^{\mu_m}$$

be the pattern of the sequence obtained by removing all appearances of $i$ from $\bar{\psi}$. More generally for any set $S \subseteq [m]$ let $\bar{\psi}_S$ be the pattern obtained by omitting the symbols in $S$ from $\bar{\psi}$. Also for all $j \geq 1$, let

$$P_j(\bar{\psi}) \overset{\text{def}}{=} P\{\bar{x} = x_1x_2\ldots x_n : \psi(\bar{x}) = \bar{\psi}, j \notin \{x_1, x_2, \ldots, x_n\}\}$$

denote the probability of the set of all sequences with pattern $\bar{\psi}$ that do not contain the symbol $j$.

Clearly, for any $P$, $\bar{\psi}$, and $j$,

$$P(\bar{\psi}) = P_j(\bar{\psi}) + \sum_{i \in [m]} p_i^{\mu_i} P_j(\bar{\psi}_i)$$

and similarly for all $S \subseteq [m]$,

$$P(\bar{\psi}_S) = P_j(\bar{\psi}_S) + \sum_{i \in [m]-S} p_i^{\mu_i} P_j(\bar{\psi}_{S\cup\{i\}}).$$

Therefore

$$\frac{\partial P(\bar{\psi})}{\partial p_j} = \sum_{i_1 \in [m]} \mu_i p_i^{\mu_i-1} P_j(\bar{\psi}_{i_1})$$

$$= \sum_{i_1 \in [m]} \mu_i p_i^{\mu_i-1} P(\bar{\psi}_{i_1}) - \sum_{i_1 \in [m]} \mu_i p_i^{\mu_i-1} \sum_{i_2 \in [m]-\{i_1\}} P_j(\bar{\psi}_{i_1,i_2})$$

42
where we applied (30). Further applying (30) to $P_j(\bar{\psi}_{i_1, i_2})$ and recursively thereafter we obtain

$$\frac{\partial P(\bar{\psi})}{\partial p_j} = \sum_{i_1 \in [m]} \mu_i P_j^{\mu_{i_1} - 1} P(\bar{\psi}_{i_1}) - \sum_{i_1 \in [m]} \sum_{i_2 \in [m] - \{i_1\}} \mu_i P_j^{\mu_{i_1} + \mu_{i_2} - 1} P(\bar{\psi}_{i_1, i_2}) + \ldots$$

$$+ (-1)^{j-1} \sum_{i_1 \in [m]} \sum_{i_2 \in [m] - \{i_1\}} \ldots \sum_{i_{\ell-1} \in [m] - \{i_1, i_2, \ldots, i_{\ell-1}\}} \mu_i P_j^{\mu_{i_1} + \mu_{i_2} + \ldots + \mu_{i_{\ell-1}} - 1} P(\bar{\psi}_{i_1, i_2, \ldots, i_{\ell-1}}) + \ldots$$

$$+ (-1)^{m-1} \sum_{i_1 \in [m]} \sum_{i_2 \in [m] - \{i_1\}} \ldots \sum_{i_m \in [m] - \{i_1, i_2, \ldots, i_m\}} \mu_{i_1} P_j^{\mu_{i_1} + \mu_{i_2} + \ldots + \mu_{i_m} - 1} \ldots$$

$$= \sum_{S \subseteq [m]} (-1)^{|S| - 1} (|S| - 1)! \left( \sum_{\ell \in S} \mu_\ell \right) \frac{\partial^{|S| - 1} P(\bar{\psi}_S)}{\partial p^{|S| - 1}} - \lambda. \tag{31}$$

Let $\hat{P} = (\hat{p}_1, \hat{p}_2, \ldots)$ be the high-profile distribution of $\bar{\psi}$ and let the polynomial

$$Q(x) \overset{\text{def}}{=} \sum_{S \subseteq [m]} (-1)^{|S| - 1} (|S| - 1)! \left( \sum_{\ell \in S} \mu_\ell \right) x^{\sum_{\ell \in S} \mu_\ell - 1} P(\bar{\psi}_S) - \lambda.$$

From (29) and (31), for all $j$

$$Q(\hat{p}_j) = \left. \frac{\partial P(\bar{\psi})}{\partial p_j} \right|_{P = \hat{P}} - \lambda = 0,$$

i.e., all the probability values of $\hat{P}$ are positive roots of $Q(x)$.

Since $Q(x)$ is a polynomial of degree

$$\sum_{i \in [m]} \mu_i - 1 = n - 1.$$

It follows that

$$V(\hat{P}) \leq n - 1.$$  

Furthermore, by Descartes’ rule of signs, the number of positive roots of a polynomial is at most the number of sign changes in its coefficients, which in turn is less than its number of non-zero coefficients. The number of non-zero coefficients of $Q(x)$ is

$$N(\bar{\psi}) \overset{\text{def}}{=} \left\{ z \in \mathbb{Z} : \exists S \subseteq [m], \sum_{i \in S} \mu_i = z \right\}$$

the number of integers that equal the sum of some subset of multiplicities of $\bar{\psi}$. Therefore

$$V(\hat{P}) \leq N(\bar{\psi}).$$

The number of subsets of $[m]$ is a simple upper bound on $N(\bar{\psi})$ and hence

$$V(\hat{P}) \leq 2^m. \quad \square$$

The theorem implies that the number of distinct probabilities in any high-profile distribution is finite. Since any non-zero probability can appear only a finite number of times we obtain the following.

**Corollary 24** The discrete size $\hat{k}$ of any high-profile distribution is finite. \quad \square
14.3 Preliminary notation and identities

Before we derive upper and lower bounds on \( \hat{t} \), we define some notation that we will use in the coming sections. Let \( P = (p_1, p_2, \ldots, p_k) \) denote a discrete distribution over \([k]\) and let \( \overline{\psi} = 1^{\mu_1}2^{\mu_2}\cdots m^{\mu_m} \) denote a canonical pattern. For any \( 1 \leq \ell \leq m \), \( \overline{\psi}_\ell \) denotes the pattern resulting from the deletion of the integer \( \ell \) from \( \overline{\psi} \). For example, \( \overline{\psi}_{-2} = 1^{\mu_1}2^{\mu_2}3^{\mu_3}\cdots m - 1^{\mu_m} \).

For any \( S_1, S_2 \subset [k] \), let
\[
P_{S_1, S_2}(\overline{\psi}) = \sum_{\substack{\mathbf{\tau} : \mathbf{\psi}(\mathbf{\tau}) = \overline{\psi}, \mathbf{\tau} \subseteq A(\mathbf{\tau}), S_1 \cap A(\mathbf{\tau}) = \emptyset}} P(\mathbf{\tau})
\]
denote the probability of all sequences with pattern \( \overline{\psi} \) that contain every symbol in \( S_1 \) and no symbol from \( S_2 \). For example, if \( \overline{\psi} = 1^{\mu_1}2^{\mu_2}, k = 4, S_1 = \{1\} \), and \( S_2 = \{3\} \), then
\[
P_{S_1, S_2}(\overline{\psi}) = p_1^{\mu_1}p_2^{\mu_2} + p_1^{\mu_1}p_4^{\mu_2} + p_2^{\mu_1}p_4^{\mu_2} + p_4^{\mu_1}p_4^{\mu_2}.
\]

When one of the sets \( S_1 \) or \( S_2 \) is empty, we simply omit it from the notation, e.g., \( P_{S_1}(\overline{\psi}) \) and \( P_{-S_2}(\overline{\psi}) \). When one of the sets, say \( S_1 \) is a singleton \( \{j\} \), we sometimes use \( P_{j, -S_2}(\overline{\psi}) \) to refer to \( P_{\{j\}, -S_2}(\overline{\psi}) \).

Given this notation, the following identities are true. For any \( 1 \leq i \leq [k] \),
\[
P_i(\overline{\psi}) = \sum_{\ell=1}^{m} p_i^{\mu_\ell} P_{-i}(\overline{\psi}_\ell).
\]

and more generally for any non-intersecting sets \( S_1, S_2 \subset [k], i \notin S_1, S_2 \),
\[
P_{i, j S_1, -S_2}(\overline{\psi}) = \sum_{\ell=1}^{m} p_i^{\mu_\ell} P_{S_1, -S_2 \cup \{j\}}(\overline{\psi}_\ell).
\]

For any \( 1 \leq \ell \leq m \),
\[
P(\overline{\psi}) = \sum_{i=1}^{k} p_i^{\mu_\ell} P_{-i}(\overline{\psi}_\ell)
\]
and more generally for any set \( S \subset [k] \), and any \( 1 \leq \ell \leq m \)
\[
P_{-S}(\overline{\psi}) = \sum_{i \in [k] \setminus S} p_i^{\mu_\ell} P_{-S \cup \{i\}}(\overline{\psi}_\ell).
\]

14.4 Upper Bound on \( \hat{t} \)

In this section, we derive an upper bound on \( \hat{t} \), the support size of the high profile distribution, and derive sufficient conditions for \( \hat{t} \) to equal \( m \), the number of distinct symbols in the pattern. The latter is interesting because it characterizes the observed data for which the high profile distribution, when used as an estimator, suggests that there are no unseen symbols beyond what was observed.

Given a distribution with finite \( k \) we merge the two least likely symbols and compare the probability assigned to a pattern with that assigned by the new distribution. This technique helps us obtain an upper bound on \( \hat{t} \) in Theorem 26 and we use this bound in Corollary 27 to show that if every symbol in a pattern \( \overline{\psi} \) appears more than \( \log (m + 1) \) times then \( \hat{t} = m \).

Before we derive an upper bound on Theorem 26 we need the following lemma on the probability of having observed the least likely symbol of a distribution given the observed pattern.
Consider a finite discrete distribution $P$ over $[k]$. Without loss of generality we assume that it is monotone i.e., $p_1 \geq p_2 \geq \ldots \geq p_k$. Let $m$ be the number of symbols appearing in a pattern $\bar{\psi}$. Recall that $\bar{X} = X_1X_2\ldots X_n$ is the observed sequence and that $\mathcal{A}(\bar{X}) = \{X_1, \ldots, X_n\}$ is the the set of distinct symbols observed.

**Lemma 25** For any discrete distribution $P \in P_M$ with finite support size $k$, and any pattern $\bar{\psi}$,

$$P_k(\bar{\psi}) \leq \frac{m}{k}P(\bar{\psi}).$$

More generally, for any $W \subseteq [k]$, if $w$ is the largest element in $[k]/W$ then,

$$P_{w,-W}(\bar{\psi}) \leq \frac{m}{k-|W|}P_w(\bar{\psi}).$$

**Proof** We prove the first claim; the second follows by similar arguments. For any $X$, note that

$$m1(\psi(X) = \bar{\psi}) = \sum_{i \in [k]} 1\{i \in X, \psi(X) = \bar{\psi}\},$$

where $1(\cdot)$ is the indicator function. Taking expectation

$$mP(\bar{\psi}) = \sum_{i \in [k]} P_i(\bar{\psi}).$$

(36)

We will prove that for all $i \in [k]$,

$$P_i(\bar{\psi}) \geq P_k(\bar{\psi})$$

and the lemma follows from (36).

Observe that for all $i \in [k - 1]$,

$$P_i(\bar{\psi}) = P_{i,k}(\bar{\psi}) + P_{i,-k}(\bar{\psi})$$

$$\geq P_{i,k}(\bar{\psi}) + P_{k,-i}(\bar{\psi})$$

$$= P_k(\bar{\psi})$$

where the inequality holds because by replacing $i$ with $k$, every sequence $\bar{x}$ where $i$ appears and $k$ does not can be uniquely mapped to another sequence $\bar{x}'$, with the same pattern, where $k$ appears and $i$ does not. \qed

Using Lemma 25 in Theorem 26, we derive an upper bound on the support size of high profile distributions. For any distribution with finite discrete size $k$ we construct a new distribution with support size $k - 1$ and show that the latter assigns a larger probability to $\bar{\psi}$ if $k$ violates an upper bound. Recall that $\mu_{\text{min}}$ is the minimum multiplicity of the symbols in a pattern.

**Theorem 26** For all non-trivial profiles,

$$\hat{t} \leq m + \frac{m - 1}{2\mu_{\text{min}} - 2}.$$

**Proof** Without loss of generality $\bar{\psi}$ is canonical, namely $1^{\mu_1}2^{\mu_2}\ldots m^{\mu_m}$ where $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_m$, and $\hat{P}$ is monotone. If $\mu_{\text{min}} = 1$, the result holds trivially, hence we consider patterns with $\mu_{\text{min}} \geq 2$. By Corollary 20, $\hat{P}$ is then discrete and hence $\hat{t}$ equals $k$, the discrete size which is finite. Let $\hat{P} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k)$. Let $Q = (q_1, \ldots, q_{k-1})$ with support set $[k - 1]$ be such that for all $i \in [k - 2], q_i = \hat{p}_i$, and $q_{k-1} = \hat{p}_k + \hat{p}_{k-1}$. We will show that if

$$\hat{t} = k > m + \frac{m - 1}{2\mu_{\text{min}} - 2}.$$
then \( \hat{P}(\bar{\psi}) < Q(\bar{\psi}) \).

Observe that for all patterns \( \bar{\psi} \)

\[
\hat{P}_{\{k-1\},k}(\bar{\psi}) = Q_{\{k-1\}}(\bar{\psi}).
\]

(37)

Then

\[
\hat{P}(\bar{\psi}) = \hat{P}_{\{k-1\},k}(\bar{\psi}) + \hat{P}_{\{k\},\{k-1\}}(\bar{\psi}) + \hat{P}_{\{k\},\{k-1\}}(\bar{\psi}) + \hat{P}_{\{k-1\},\{k\}}(\bar{\psi})
\]

\[
= Q_{\{k-1\}}(\bar{\psi}) + \hat{P}_{\{k\},\{k-1\}}(\bar{\psi}) + \hat{P}_{\{k-1\},\{k\}}(\bar{\psi}) = Q_{\{k-1\}}(\bar{\psi}) + \sum_{i=1}^{m} \hat{P}_{\{k\},\{k-1\}}(\bar{\psi}) + \sum_{i=1}^{m} \hat{P}_{\{k-1\},\{k\}}(\bar{\psi})
\]

\[
= Q_{\{k-1\}}(\bar{\psi}) + \sum_{i=1}^{m} \hat{P}_{\{k\},\{k-1\}}(\bar{\psi}) + \sum_{i=1}^{m} \hat{P}_{\{k-1\},\{k\}}(\bar{\psi})
\]

(38)

where the second and third equalities follow from (37).

From Lemma 25,

\[
\hat{P}_{\{k-1\},\{k\}}(\bar{\psi}) \leq \frac{m-1}{k-1} \hat{P}_{\{k\}}(\bar{\psi}).
\]

But note that

\[
\hat{P}_{\{k\}}(\bar{\psi}) = \hat{P}_{\{k\},\{k-1\}}(\bar{\psi}) + \hat{P}_{\{k\},\{k-1\}}(\bar{\psi})
\]

and therefore

\[
\hat{P}_{\{k\},\{k-1\}}(\bar{\psi}) \leq \frac{m-1}{k-1} \hat{P}_{\{k\}}(\bar{\psi}).
\]

Substituting in (38) and combining with (37) we obtain that

\[
\hat{P}(\bar{\psi}) \leq Q_{\{k-1\}}(\bar{\psi}) + \sum_{i=1}^{m} \left( \hat{P}_{\{k\},\{k-1\}} + \hat{P}_{\{k-1\},\{k\}} \right) Q_{\{k-1\}}(\bar{\psi})
\]

(39)

But

\[
Q(\bar{\psi}) = Q_{\{k-1\}}(\bar{\psi}) + Q_{\{k\}}(\bar{\psi})
\]

\[
= Q_{\{k-1\}}(\bar{\psi}) + \sum_{i=1}^{m} q_{i\{k\}} \hat{P}_{\{k\},\{k-1\}}(\bar{\psi})
\]

\[
\stackrel{(a)}{=} Q_{\{k-1\}}(\bar{\psi}) + \sum_{i=1}^{m} (\hat{p}_{k} + \hat{p}_{k-1})^{\mu_{i}} Q_{\{k\}}(\bar{\psi})
\]

\[
= Q_{\{k-1\}}(\bar{\psi}) + \sum_{i=1}^{m} \left( \hat{p}_{k}^{\mu_{i}} + \hat{p}_{k-1}^{\mu_{i}} + \sum_{a=1}^{\mu_{i}-1} \left( \frac{\mu_{i}}{a} \right) \frac{\hat{p}_{k-1}^{\mu_{i}-a}}{\hat{p}_{k}^{\mu_{i}}} \right) Q_{\{k\}}(\bar{\psi})
\]

\[
\stackrel{(b)}{=} Q_{\{k-1\}}(\bar{\psi}) + \sum_{i=1}^{m} \left( \hat{p}_{k}^{\mu_{i}} + \hat{p}_{k-1}^{\mu_{i}} + \sum_{a=1}^{\mu_{i}-1} \left( \frac{\mu_{i}}{a} \right) \hat{p}_{k}^{\mu_{i}} \right) Q_{\{k\}}(\bar{\psi})
\]

where (a) follows from the construction of \( Q \) and (b) because \( \hat{p}_{k-1} \geq \hat{p}_{k} \). Subtracting this from (39) we obtain

\[
\hat{P}(\bar{\psi}) - Q(\bar{\psi}) \leq \sum_{i=1}^{m} \left( \frac{m-1}{k-m} - \sum_{a=1}^{\mu_{i}-1} \left( \frac{\mu_{i}}{a} \right) \right) \hat{p}_{k}^{\mu_{i}} Q_{\{k\}}(\bar{\psi})
\]

(40)

For all \( i \),

\[
\sum_{a=1}^{\mu_{i}-1} \left( \frac{\mu_{i}}{a} \right) = 2^{\mu_{i}} - 2.
\]
Note that, since $Q_{\{k-1\}}(\tilde{\psi}_{-i}) = \hat{P}_{\{k-1,k\}}(\tilde{\psi}_{-i})$ and the latter is positive for all non-trivial patterns $\tilde{\psi}$ and all $i$, if for all $i$

$$\frac{m-1}{k-m} - (2^{\mu_i} - 2) < 0,$$

then $Q(\tilde{\psi}) > \hat{P}(\tilde{\psi})$, a contradiction. Therefore there exists at least one $i$ for which

$$\frac{m-1}{k-m} \geq 2^{\mu_i} - 2.$$

Since $\mu_{\text{min}}$ is the minimum multiplicity we have

$$k \leq m + \frac{m-1}{2^{\mu_{\text{min}}} - 2} \quad \Box$$

We use the upper bound of Theorem 26 to show that for any pattern $\tilde{\psi}$ with $m$ symbols if every symbol appears more than $\log(m + 1)$ times then the support size of any high-profile distribution equals $m$. This corollary reduces the search for the high profile distributions of a large class of patterns to performing a maximization over $m$ dimensions, where $m$ is the number of distinct symbols in the pattern.

**Corollary 27** For all non-trivial patterns such that $\mu_{\text{min}} > \log(m + 1)$,

$$\hat{t} = m.$$

**Proof** Clearly $\hat{t} \geq m$ and from Theorem 26,

$$\hat{t} \leq m + \frac{m-1}{2^{\mu_{\text{min}}} - 2} < m + 1. \quad \Box$$

### 14.5 Lower bound on $\hat{t}$

In this section we lower bound the support size of general and some specific profiles.

In Lemma 28 we provide a simple proof that for the uniform profile $m < \hat{t} < \infty$ namely, it is finite but strictly larger than the number of observed symbols. This implies that one can find arbitrarily long sequences for which $m < \hat{t} < \infty$.

In Theorem 30 we derive a general lower bound on $\hat{t}$, and use it in Corollary 32 to show that there are profiles with $\alpha m < \hat{t} < \infty$, for some $\alpha > 1$ and in Corollary 31 where we prove a complementary result to Corollary 27. While the latter showed that if $\mu_{\text{min}} > \log(m + 1)$, $\hat{t} = m$, Corollary 31 shows that if $\mu_{\text{max}} < \log(\sqrt{m} + 1)$, $\hat{t} > m$. This is interesting because it characterizes the observed data for which the high profile distribution, when used as an estimator, suggests that there are unseen symbols beyond what was observed.

**Lemma 28** For all $r \geq 2$ and all sufficiently large $m$

$$m < t_{r,m} < \infty.$$

**Proof** Since all symbols have multiplicity $r$, Theorem 22 implies that $t_{r,m}$ is finite. Consider a distribution $P$ over $[m]$. Then

$$P(\tilde{\psi}) = m! \prod_{i=1}^{m} p_i^r.$$

47
Therefore, the distribution with support size $m$ that maximizes $P(\bar{\psi})$ is uniform. It assigns a probability $\frac{m!}{m^m}$ to the pattern. Let $P'$ be the uniform distribution over a set of $m+1$ elements. It can also be verified that $P(\bar{\psi}) = \frac{(m+1)!}{(m+1)^m}$. For all sufficiently large $m$,

$$\frac{P'(\bar{\psi})}{P(\bar{\psi})} = \frac{m+1}{(1 + 1/m)^m} > 1,$$

hence $m < t_{r,m} < \infty$. \hfill \Box

There are therefore arbitrarily long patterns $1^r 2^r \ldots m^r$, whose high-profile support is larger than the number of observed symbols. Later in this section we show that $\hat{t}$ is at least $\alpha m$ where $\alpha$ is a constant that depends on $r$.

But first we lower bound the high-profile support of general profiles. As in Section 14.4, we begin with preliminary results.

Consider a finite discrete distribution $P$ over $[k]$. Without loss of generality we assume that it is monotone i.e., $p_1 \geq p_2 \geq \ldots \geq p_k$. Let $m$ be the number of symbols appearing in a pattern $\bar{\psi}$. Recall that $\mathbf{X} = X_1 X_2 \ldots X_n$ is the sequence of observed elements and that $\mathcal{A}(\mathbf{X}) = \{X_1, \ldots, X_n\}$ is the the set of distinct symbols observed.

**Lemma 29** Let $P$ be a discrete distribution over $[k]$ where $p_j \geq c p_i$ for all $i \in [k]$. Then for every pattern $\bar{\psi}$,

$$P_j(\bar{\psi}) \geq \frac{\sum_{i=1}^{m} c^{\ell_i}}{k - m + \sum_{i=1}^{m} c^{\ell_i}} P(\bar{\psi}).$$

More generally, if $W \subseteq [k]$ and $j \notin W$ is such that $p_j \geq c p_i$ for all $i \in [k]/W$, then

$$P_{j,W}(\bar{\psi}) \geq \frac{\sum_{i=1}^{m} c^{\ell_i}}{k - |W| - m + \sum_{i=1}^{m} c^{\ell_i}} P_{W}(\bar{\psi}).$$

**Proof** We prove the first claim, the second follows similarly. The result is trivially true if $k = m$. Therefore we consider $k > m$. Observe that for all $j \in [k]$

$$1(j \in \mathcal{A}(\mathbf{X}), \psi(\mathbf{X}) = \bar{\psi}) = \frac{1}{k-m} \sum_{i=1}^{k} 1(j \in \mathcal{A}(\mathbf{X}), i \notin \mathcal{A}(\mathbf{X}), \psi(\mathbf{X}) = \bar{\psi})$$

since the number of symbols appearing in $\bar{\psi}$ is $m$. Taking expectation on both sides we obtain that

$$P_j(\bar{\psi}) = \frac{1}{k-m} \sum_{i=1, i \neq j}^{k} P_{j-i}(\bar{\psi})$$

$$= \frac{1}{k-m} \sum_{i=1, i \neq j}^{k} \sum_{\ell=1}^{m} c^{\ell_i} p_j^{\ell_i} P_{-(i,j)}(\bar{\psi}_{-\ell})$$

$$\geq \frac{1}{k-m} \sum_{i=1, i \neq j}^{k} \sum_{\ell=1}^{m} c^{\ell_i} p_j^{\ell_i} P_{-(i,j)}(\bar{\psi}_{-\ell})$$

$$= \frac{1}{k-m} \sum_{\ell=1}^{m} c^{\ell_i} \sum_{i=1, i \neq j}^{k} p_j^{\ell_i} P_{-(i,j)}(\bar{\psi}_{-\ell})$$

$$= \frac{1}{k-m} \sum_{\ell=1}^{m} c^{\ell_i} P_j(\bar{\psi})$$

48
where the second equality follows from (33), the inequality from the fact that for all \( i, p_j \geq c p_i \), and the last equality from (35). Since \( P(\hat{p} r t) = P_j(\hat{p}) + P_{-j}(\hat{p}) \), it follows that
\[
P_j(\hat{p}) \geq \frac{\sum_{i=1}^{m} c_{ij}}{k - m + \sum_{i=1}^{m} c_{ii}} P(\hat{p}). \tag*{\Box}
\]

In Theorem 30 we derive a lower bound on the high-profile support. We do so by proceeding along the lines of the proof of Theorem 26 where for any distribution whose support size was larger than \( m \) we constructed a new distribution of smaller size. However in this theorem, the constructed distribution’s support size is one larger than the one we began with. If the original distribution was indeed the best distribution, then it imposes a lower bound on its support size. Recall that \( \mu_{\text{max}} \) is the maximum multiplicity of a pattern.

**Theorem 30** For all non-trivial profiles,
\[
\hat{t} \geq m - 1 + \left( \sum_{j \in [m]} 2^{-\mu_j} - 2^{-\mu_{\text{max}}} \right).
\]

**Proof** Without loss of generality \( \bar{\psi} = 1^{\mu_1} \bar{\mu}_2 \ldots m^{\mu_m} \) is canonical, namely \( \mu_1 \geq \mu_2 \geq \ldots \geq \mu_m \) and \( \bar{P} = (\bar{p}_1, \bar{p}_2, \ldots) \) is monotone. If \( \hat{t} \) is finite the bound holds trivially. Therefore we need to only consider the case when \( \hat{t} \) is infinite the bound holds trivially. We construct a new distribution \( Q = (q_1, \ldots, q_{k+1}) \) over \([k + 1]\) such that for all \( 2 \leq i \leq k, q_i = \bar{p}_i \), and \( q_1 = q_{k+1} = \bar{p}_1/2 \), and show that if
\[
\hat{t} = k < m - 1 + \left( \sum_{j \in [m]} 2^{-\mu_j} - 2^{-\mu_{\text{max}}} \right),
\]
then \( \bar{P}(\bar{\psi}) < Q(\bar{\psi}) \).

Observe that
\[
Q(\bar{\psi}) = Q_{-\{1,k+1\}}(\bar{\psi}) + Q_{-\{1\}}(\bar{\psi}) + Q_{k+1,-1}(\bar{\psi}) + Q_{\{1,k+1\}}(\bar{\psi}) \]
\[
= Q_{-\{1,k+1\}}(\bar{\psi}) + \sum_{\ell=1}^{m} q_{\ell}\bar{\mu}_1 Q_{-\{1,k+1\}}(\bar{\psi} - \ell) + \sum_{\ell=1}^{m} q_{\ell}\bar{\mu}_1 Q_{\{1,k+1\}}(\bar{\psi} - \ell) + \sum_{\ell=1}^{m} q_{\ell}\bar{\mu}_1 Q_{k+1,-1}(\bar{\psi} - \ell) \tag{41}
\]
where the second equality follows from (33). By construction, for all \( i \in \{2, \ldots, k\} \),
\[
q_{k+1} = \frac{\bar{p}_1}{2} \geq \frac{\bar{p}_i}{2} = q_i.
\]

Applying Lemma 29 with the pattern \( \bar{\psi} \), \( c = \frac{1}{2} \) and \( W = \{1\} \), we obtain that
\[
Q_{k+1,-1}(\bar{\psi} - \ell) \geq \frac{\sum_{i=1, i \neq \ell}^{m} 2^{-\mu_i}}{k - m + 1 + \sum_{i=1, i \neq \ell}^{m} 2^{-\mu_i}} Q_{-\{1\}}(\bar{\psi} - \ell).
\]

Since
\[
Q_{-1}(\bar{\psi} - \ell) = Q_{k+1,-1}(\bar{\psi} - \ell) + Q_{-\{1,k+1\}}(\bar{\psi} - \ell)
\]
it follows that
\[
Q_{k+1,-1}(\bar{\psi} - \ell) \geq \frac{\sum_{i=1, i \neq \ell}^{m} 2^{-\mu_i}}{k - m + 1} Q_{-\{1,k+1\}}(\bar{\psi} - \ell).
\]
Substituting in (41), we obtain that
\[ Q(\bar{\psi}) \geq Q_{-\{1,k+1\}}(\bar{\psi}) + \sum_{\ell=1}^{m} \left( q_{\ell}^{\mu_{\ell}} + q_{k+1}^{\mu_{k}} + q_{1}^{\mu_{1}} \sum_{i=1,i\neq \ell}^{m} \frac{2^{-\mu_{i}}}{k-m+1} \right) Q_{-\{1,k+1\}}(\bar{\psi}-\ell) \]
\[ = \hat{P}_{-1}(\bar{\psi}) + \sum_{\ell=1}^{m} \left( \frac{\hat{p}_{\ell}}{2} \right)^{\mu_{\ell}} + \left( \frac{\hat{p}_{1}}{2} \right)^{\mu_{1}} + \left( \frac{\hat{p}_{k+1}}{2} \right)^{\mu_{k+1}} \sum_{i=1,i\neq \ell}^{m} \frac{2^{-\mu_{i}}}{k-m+1} \right) \hat{P}_{-1}(\bar{\psi}-\ell) \]
\[ = \hat{P}_{-1}(\bar{\psi}) + \sum_{\ell=1}^{m} \left( 2^{-(\mu_{\ell}-1)} + 2^{-\mu_{1}} \sum_{i=1,i\neq \ell}^{m} \frac{2^{-\mu_{i}}}{k-m+1} \right) \hat{P}_{-1}(\bar{\psi}-\ell) \quad (42) \]
where the first equality follows from the construction of \( Q \). If for all \( 1 \leq \ell \leq m \),
\[ 2^{-(\mu_{\ell}-1)} + 2^{-\mu_{1}} \sum_{i=1,i\neq \ell}^{m} \frac{2^{-\mu_{i}}}{k-m+1} > 1 \]
then (42) reduces to
\[ Q(\bar{\psi}) > \hat{P}_{-1}(\bar{\psi}) + \sum_{\ell=1}^{m} \hat{p}_{\ell}^{\mu_{\ell}} \hat{P}_{-1}(\bar{\psi}-\ell) \]
\[ = \hat{P}_{-1}(\bar{\psi}) + \hat{P}_{1}(\bar{\psi}) \]
\[ = \hat{P}(\bar{\psi}) \]
where the first equality follows from (32). But this is a contradiction. Therefore there exists at least one \( \ell \in [m] \) such that
\[ 2^{-(\mu_{\ell}-1)} + 2^{-\mu_{1}} \sum_{i=1,i\neq \ell}^{m} \frac{2^{-\mu_{i}}}{k-m+1} \leq 1 \]
or equivalently there exists at least one \( \ell \in [m] \) such that
\[ k \geq m - 1 + \frac{\sum_{i=1,i\neq \ell}^{m} 2^{-\mu_{i}}}{2^{\mu_{\ell}} - 2} \]
namely,
\[ k \geq m - 1 + \min_{\ell} \left( \frac{\sum_{i=1,i\neq \ell}^{m} 2^{-\mu_{i}}}{2^{\mu_{\ell}} - 2} \right) \].
We will show that this minimum is achieved when \( \mu_{\ell} = \mu_{\max} \).

If \( \mu_{\ell_{1}} = \mu_{\ell_{2}} \) for all \( \ell_{1} \neq \ell_{2} \), then we have the required result. If not, let \( \mu_{\ell_{1}} > \mu_{\ell_{2}} \) for some \( \ell_{1}, \ell_{2} \). We will show that
\[ \frac{\sum_{i=1,i\neq \ell}^{m} 2^{-\mu_{i}}}{2^{\mu_{\ell_{1}} - 2}} < \frac{\sum_{i=1,i\neq \ell}^{m} 2^{-\mu_{i}}}{2^{\mu_{\ell_{2}} - 2}} \quad (43) \]
which proves that
\[ \min_{\ell} \left( \frac{\sum_{i=1,i\neq \ell}^{m} 2^{-\mu_{i}}}{2^{\mu_{\ell}} - 2} \right) = \left( \frac{\sum_{j \in [m]} 2^{-\mu_{j}} - 2^{-\mu_{\max}}}{2^{\mu_{\max}} - 2} \right) \).

Let
\[ c \overset{\text{def}}{=} \sum_{i=1}^{m} 2^{-\mu_{i}} \).

If \( \mu_{\ell_{2}} = 1 \), then (43) holds trivially. Otherwise we need to show that
\[ (c - 2^{-\mu_{\ell_{1}}})(2^{\mu_{\ell_{2}} - 2}) - (c - 2^{-\mu_{\ell_{2}}})(2^{\mu_{\ell_{1}} - 2}) < 0. \]
Observe that

\[
(c - 2^{-\mu_{\ell_1}})(2^{\mu_{\ell_2}} - 2) - (c - 2^{-\mu_{\ell_2}})(2^{\mu_{\ell_1}} - 2) \\
= 2(\mu_{\ell_1} - \mu_{\ell_2}) - 2(-2^{-\mu_{\ell_2}} - 2^{-\mu_{\ell_1}}) - c(2^{\mu_{\ell_1}} - 2^{\mu_{\ell_2}}),
\]

\[
= \frac{(2^{\mu_{\ell_1}} + 2^{\mu_{\ell_2}})(2^{\mu_{\ell_1}} - 2^{\mu_{\ell_2}})}{2^{\mu_{\ell_1} + \mu_{\ell_2}}} - 2^{\mu_{\ell_1}} - 2^{\mu_{\ell_2}} - c(2^{\mu_{\ell_1}} - 2^{\mu_{\ell_2}}),
\]

\[
= (2^{\mu_{\ell_1}} - 2^{\mu_{\ell_2}})(2^{-\mu_{\ell_1}} + 2^{-\mu_{\ell_2}} - 2(2^{-\mu_{\ell_1} + \mu_{\ell_2}}) - c).
\]

As \(\mu_{\ell_1} > \mu_{\ell_2}\), we have \((2^{\mu_{\ell_1}} - 2^{\mu_{\ell_2}}) > 0\) and from the definition of \(c\),

\[c \geq 2^{-\mu_{\ell_1}} + 2^{-\mu_{\ell_2}}.\]

Therefore we have the required result. \(\square\)

Using this theorem one can prove several results about \(\hat{t}\). In the following corollary, we derive a sufficient condition for \(\hat{t}\) to be larger than the number of observed symbols.

**Corollary 31** For all non-trivial profiles, if \(\mu_{\text{max}} < \log (\sqrt{m} + 1)\)

\[\hat{t} > m.\]

**Proof** From Theorem 30 if

\[\sum_{i \in [m]} 2^{-\mu_i} - 2^{-\mu_{\text{max}}} > 2^{\mu_{\text{max}}} - 2\]

then

\[\hat{t} > m.\]

Since \(\mu_{\text{max}} < \log (\sqrt{m} + 1)\),

\[\sum_{i \in [m]} 2^{-\mu_i} - 2^{-\mu_{\text{max}}} \geq (m - 1)2^{-\mu_{\text{max}}} > (m - 1) \frac{1}{\sqrt{m} + 1} \geq \sqrt{m} - 1 > 2^{\mu_{\text{max}}} - 2. \]

\(\square\)

Theorem 30 also helps us address the question of how large the support set of \(\hat{P}\) can be. In the following corollary we consider the uniform profile \(r^m\) and show that for all \(r \geq 2\) and all sufficiently large \(m\), \(\hat{t} = \alpha m\), where \(\alpha > 1\) is a constant that depends on \(r\).

**Corollary 32** For all \(r \geq 2\),

\[t_r \geq \alpha m,\]

where \(\alpha > 1\).

**Proof** Since \(r \geq 2\), \(\hat{t}\) is finite. From Theorem 30,

\[\hat{t} \geq m - 1 + \left(\frac{\sum_{j \in [m]} 2^{-\mu_j} - 2^{-\mu_{\text{max}}}}{2^{\mu_{\text{max}}} - 2}\right) = (m - 1) \left(1 + \frac{2^{-r}}{2^{r} - 2}\right). \]

\(\square\)

In Theorem 45 we will show that \(\hat{P}_r\) is the uniform distribution over \([\hat{k}]\), where

\[\hat{k} = \min \left\{ k \geq m : \left(1 + \frac{1}{k}\right)^{kr} \left(1 - \frac{m}{k + 1}\right) > 1 \right\}.\]

51
14.6 Partial order on patterns

We define a partial order on patterns according to their probabilities under any distribution, and hence also their PML. We use this order to find the patterns with highest and lowest PML.

Let \( \pi = x_1, \ldots, x_m \) and \( \mu = \mu_1, \ldots, \mu_m \) be two \( m \)-valued real sequences. The \( \pi \) power of \( \pi \) is

\[
\pi^\pi \overset{\text{def}}{=} \prod_{i=1}^{m} x_i^{\mu_i},
\]

the product of all corresponding powers. The symmetric \( \pi \) power of \( \pi \) is

\[
\left[ \pi^\pi \right] \overset{\text{def}}{=} \sum_{\sigma \in [m]^m} \pi^{(\mu_{\sigma(1)}, \ldots, \mu_{\sigma(m)})},
\]

the sum of \( \pi \) raised to \( \mu \) and all its \( m! \) index permutations. For example,

\[
\left[ (x, y)^{(1, 2)} \right] = (x, y)^{(1, 2)} + (x, y)^{(2, 1)} = xy^2 + x^2y,
\]

and

\[
\left[ (x, y)^{(1, 1)} \right] = (x, y)^{(1, 1)} + (x, y)^{(1, 1)} = 2xy.
\]

Let \( \mu = (\mu_1, \ldots, \mu_m) \) and \( \mu' = (\mu'_1, \ldots, \mu'_m) \) be two non-increasing, non-negative, real-number sequences of the same length \( m \) and sum

\[
\sum_{i=1}^{m} \mu_i = \sum_{i=1}^{m} \mu'_i.
\]

\( \mu \) majorizes \( \mu' \), written \( \mu \succeq \mu' \), if for all \( 1 \leq i < m \),

\[
\sum_{j=1}^{i} \mu_j \geq \sum_{j=1}^{i} \mu'_j.
\]

For example, \( (5, 2, 1) \succeq (3, 3, 2) \) as both are non-increasing, non-negative, 3-element sequences summing to 8, and \( 5 \geq 3 \) and \( 5 + 2 \geq 3 + 3 \).

Muirhead’s Inequality [7] says that if \( \mu \succeq \mu' \) then for all nonnegative vectors \( \pi \) of the same length,

\[
\left[ \pi^\pi \right] \geq \left[ \pi'^\pi' \right].
\]

For example, \( (3, 0) \succeq (2, 1) \), hence for all nonnegative \( x \) and \( y \),

\[
x^3 + y^3 \geq x^2y + y^2x,
\]

as can be verified by considering \( x \geq y \) and \( x \leq y \). Similarly, \( (1, 0, \ldots, 0) \succeq (1/m, \ldots, 1/m) \), hence for all nonnegative \( x_1, \ldots, x_m \),

\[
\left[ (x_1, \ldots, x_m)^{(1, 0, \ldots, 0)} \right] \geq \left[ (x_1, \ldots, x_m)^{(1/m, \ldots, 1/m)} \right].
\]

Evaluating both sides and normalizing by \( m! \) we obtain the arithmetic-geometric mean inequality

\[
\frac{1}{m} \sum_{i=1}^{m} x_i \geq \left( \prod_{i=1}^{m} x_i \right)^{1/m}.
\]

Let \( \bar{\psi} \) and \( \bar{\psi}' \) be two canonical patterns of length \( sn \) and \( m \) symbols. \( \bar{\psi} \) majorizes \( \bar{\psi}' \), written \( \bar{\psi} \succeq \bar{\psi}' \), if its multiplicity sequence \( \mu_1, \ldots, \mu_m \) majorizes that of \( \bar{\psi}' \). For example, the canonical
pattern $1^52^31^3$ majorizes the canonical pattern $1^32^33^2$ because as we saw, $(5, 2, 1) \succeq (3, 3, 2)$. If the patterns $\tilde{\psi}$ and $\tilde{\psi}'$ are of the same length and number of symbols, but are not necessarily canonical, we say that $\tilde{\psi}$ majorizes $\tilde{\psi}'$ if the canonical pattern of $\tilde{\psi}$ majorizes that of $\tilde{\psi}'$. For example, $1^22^13^5$ whose canonical pattern is $1^52^31^3$ majorizes $1^22^33^3$ whose canonical pattern is $1^32^33^2$. Note that majorization is defined only for patterns of the same length and number of symbols.

Muirhead’s Inequality implies that pattern probabilities are ordered by majorization.

**Theorem 33** If $\tilde{\psi} \succeq \tilde{\psi}'$ then for every distribution $P$,

$$ P(\tilde{\psi}) \geq P(\tilde{\psi}') . $$

**Proof** If $P$ is discrete, the theorem is exactly Muirhead’s Inequality,

$$ P(\tilde{\psi}) = \sum_{S \in \binom{[n]}{k}} [(P_S)_{\tilde{\psi}}] \geq \sum_{S \in \binom{[n]}{k}} [(P_S)_{\tilde{\psi}'}] = P(\tilde{\psi}') , $$

where $k$ can be finite or infinite. When $P$ has a continuous part, the proof is similar. Since $\tilde{\psi} \succeq \tilde{\psi}'$, necessarily $\varphi_1 \geq \varphi_1'$, hence, assuming canonical form,

$$ P(\tilde{\psi}) = \sum_{i=0}^{\varphi_1} \binom{\varphi_1}{i} q^i P_{-q}(\psi_1 \ldots \psi_{n-i}) $$

$$ \geq \sum_{i=0}^{\varphi_1'} \binom{\varphi_1'}{i} q^i P_{-q}(\psi_1' \ldots \psi_{n-i}') $$

$$ \geq \sum_{i=0}^{\varphi_1'} \binom{\varphi_1'}{i} q^i P_{-q}(\psi_1' \ldots \psi_{n-i}') $$

$$ = P(\tilde{\psi}'). \quad \square $$

For example, $1^32 \succeq 1^22$, hence for all distributions, $P(1^32) \geq P(1^22)$. However, unlike Muirhead’s Inequality where the power vectors can have different number of 0’s, pattern majorization is defined, and hence the theorem holds, only for patterns with the same number of symbols. For example, although $(2, 0) \succeq (1, 1)$, the pattern $1^2$ does not majorize 12 as they have 1 and 2 symbols respectively. Indeed the continuous distribution $P = ()$ has $P(1^2) = 0 < 1 = P(12)$.

**Corollary 34** If $\tilde{\psi} \succeq \tilde{\psi}'$, then

$$ \hat{P}(\tilde{\psi}) \geq \hat{P}(\tilde{\psi}'). \quad \square $$

The corollary determines the patterns of highest and lowest PML for any given length and number of symbols. A pattern of the form $1^r2^r \ldots m^r$ is **skewed**, a pattern of the form $1^r2^r \ldots m^r$ is **uniform**, and a pattern of the form $1^{r+1} \ldots i^{r+1}(i+1)^r \ldots m^r$ is **1-uniform**. For example, $1^223$ is skewed, $1^32^33^3$ is uniform, and $1^32^33^3$ is 1-uniform.

For any given length $n$ and number of symbols $m$ there is a unique skewed pattern $\tilde{\psi}s = 1^{n-m+1}2 \ldots m$ and a unique uniform or 1-uniform pattern $\tilde{\psi}_u$. It is easy to see that all patterns $\tilde{\psi}$ with these $n$ and $m$ are majorized by $\tilde{\psi}_s$ and majorize $\tilde{\psi}_u$, hence

$$ \hat{P}(\tilde{\psi}_s) \geq \hat{P}(\tilde{\psi}) \geq \hat{P}(\tilde{\psi}_u). $$

For example, for $n = 6$ and $m = 3$, $\hat{P}(1^423) \geq \hat{P}(1^32^33) \geq \hat{P}(1^22^33^2)$.
15 Specific profiles

In the previous section we derived some general properties of high-profile distributions. Despite these results, deriving the high-profile distribution of profiles in general, appears to be a hard problem, and therefore in Section 15 we consider some interesting special profiles, whose structure makes the problem somewhat tractable. We consider four kinds of special profiles – profiles of length \( \leq 4 \), skewed profiles which comprise of one symbol appearing several times and all other symbols appearing once, quasi-uniform profiles, where the multiplicities of any two symbols differ by no more than 1, and binary profiles which contain exactly two symbols.

15.1 Short Profiles

We determine the high-profile distributions of all profiles of length \( \leq 4 \). The shortest canonical pattern with a non-trivial high-profile distribution is 112. This pattern belongs to the larger class of patterns with profile \( r \rightarrow \rightarrow \), where \( r \geq 2 \). In Theorem 35, we show that for all such patterns, \( \hat{t} = 2 \). To do so, we prove that for all distributions with support size \( > 2 \) it is possible to construct distributions with smaller support size that assign a greater probability to the pattern. The new distribution is constructed by adding the two smallest probabilities in the original distribution and retaining the other probabilities as is.

Theorem 35 For all \( r \geq 2 \),

\[ t_{r+1} = 2. \]

Proof Wolog, \( \bar{\psi} = 1^r2 \). From Theorem 21, \( \hat{t} \) is finite. Consider a discrete distribution \( P = (p_1, \ldots, p_k) \) over \( [k] \) where \( p_i \geq p_{i+1} \) for \( i \in [k-1] \). If \( k > 2 \), construct a new distribution \( Q \) of support size \( k - 1 \) where \( q_i = p_i \) for \( i \in [k-2] \), and \( q_{k-1} = p_{k-1} + p_k \). We show that \( Q(\bar{\psi}) > P(\bar{\psi}) \).

Since \( k - 1 \) and \( k \) can either appear or not,

\[ P(\bar{\psi}) = P_{\{k-1,k\}}(\bar{\psi}) + P_{k-1,-k}(\bar{\psi}) + P_{k,-(k-1)}(\bar{\psi}) + P_{\{k-1,k\}}(\bar{\psi}). \]

Observe that

\[ P_{k-1,-k}(\bar{\psi}) + P_{k,-(k-1)}(\bar{\psi}) = \sum_{i \in [k-2]} p_i^r (p_{k-1} + p_k) + (p_{k-1}^r + p_k^r) \sum_{i \in [k-2]} p_i, \]

and

\[ P_{\{k-1,k\}}(\bar{\psi}) = p_{k-1}^r p_k + p_k^r p_{k-1} \leq (p_{k-1}^{r-1} p_k + p_k^{r-1} p_{k-1}) \sum_{i \in [k-2]} p_i \]

where the inequality holds as \( p_i \geq p_{i+1} \) for \( i \in [k-1] \). Substituting (45) and (46) in (44) we
obtain that
\[
P(\bar{\psi}) \leq P_{\{k-1,k\}}(\bar{\psi}) + \sum_{i \in [k-2]} p_i^\psi (p_{k-1} + p_k) + \left( p_{k-1}^\psi + p_k^\psi \right) \sum_{i \in [k-2]} p_i + \left( p_{k-1}^{r-1}p_k + p_k^{r-1}p_{k-1} \right) \sum_{i \in [k-2]} p_i
\]
\[
= P_{\{k-1,k\}}(\bar{\psi}) + \sum_{i \in [k-2]} p_i^\psi (p_{k-1} + p_k) + \left( p_{k-1}^\psi + p_k^\psi \right) \sum_{i \in [k-2]} p_i
\]
\[
\leq P_{\{k-1,k\}}(\bar{\psi}) + \sum_{i \in [k-2]} q_i^\psi q_{k-1} + q_k^{r-1} \sum_{i \in [k-2]} q_i
\]
\[
= Q_{\{k-1\}}(\bar{\psi}) + Q_{k-1}(\bar{\psi}) = Q(\bar{\psi}) \tag{47}
\]
where the equality following the second inequality in the chain follows from the construction of \( Q \). Equalities in (47) hold only if \( r = 2, k = 3 \), and \( p_1 = p_2 = p_3 \). In that case a simple calculation shows that the distribution \((1/2, 1/2)\) assigns a greater probability to the pattern. Therefore if \( k > 2 \), \( P \) cannot be a high-profile distribution. Hence \( t = 2 \).

We apply this general result to the profile \( 2^11^1 \), the shortest profile with a non-trivial high-profile distribution.

**Corollary 36**

\[
\hat{P}_{2^11^1} = (1/2, 1/2),
\]
and for all patterns \( \bar{\psi} \in \bar{\Psi}_{2^11^1} \)
\[
\hat{P}(\bar{\psi}) = 1/4.
\]

**Proof** Wolog let \( \bar{\psi} = 112 \). From Theorem 35, \( \hat{t} = 2 \). The probability that \( P = (p, 1-p) \) assigns to \( \bar{\psi} \) is
\[
P(\bar{\psi}) = p^2(1-p) + (1-p)^2p = p(1-p),
\]
which is maximized when \( p = 1/2 \), and hence \( \hat{P} = (1/2, 1/2) \) and \( \hat{P}(\bar{\psi}) = 1/4 \).

We have therefore derived the high-profile distributions of all profiles of length at most 3 and now consider profiles of length 4. The constant profile \( \bar{\varphi} = 4^1 \) and the distinct profile \( \bar{\varphi} = 1^4 \) were addressed in part I of this paper and we now consider the profile \( 3^11^1 \).

**Corollary 37**

\[
\hat{P}_{3^11^1} = (1/2, 1/2)
\]
and for all patterns \( \bar{\psi} \in \bar{\Psi}_{3^11^1} \)
\[
\hat{P}(\bar{\psi}) = 1/8.
\]

**Proof** Wolog let \( \bar{\psi} = 1112 \). From Theorem 35, \( \hat{t} = 2 \) and the probability that \( P = (p, 1-p) \) assigns to \( \bar{\psi} \) is
\[
P(\bar{\psi}) = p^3(1-p) + (1-p)^3p.
\]
Differentiating \( P(\bar{\psi}) \) with respect to \( p \) and setting it to 0 we get
\[
3p^2(1-p) - p^3 - 3(1-p)^2p + (1-p)^3 = 1 - 6p + 12p^2 - 8p^3 = (1 - 2p)^3 = 0.
\]
Hence the only maxima is at \( p = 1/2 \). Therefore \( \hat{P} = (1/2, 1/2) \) and \( \hat{P}(\bar{\psi}) = 1/8 \).
The next shortest profile is $2^2$. While Theorem 35 does not apply to it, Corollary 27 does and we get that $\hat{t} = 2$. That $\hat{P} = (1/2, 1/2)$ follows easily.

**Theorem 38**

$$\hat{P}_{2^2} = (1/2, 1/2),$$

and for all patterns $\psi \in \Psi_{2^2}$

$$\hat{P}(\psi) = 1/8.$$

**Proof** Consider the canonical pattern $\psi = 1122$. Since

$$\mu_{\min} = 2 > \log(2 + 1) = \log(m + 1),$$

from Corollary 27, $\hat{t} = 2$. The probability that $P = (p, 1 - p)$ assigns to the pattern 1122 is

$$P(\psi) = 2p^2(1 - p)^2$$

which is maximized when $p = 1/2$, and hence $\hat{P} = (1/2, 1/2)$ and $\hat{P}(\psi) = 1/8$. \hfill \Box

The final length 4 profile is $2^11^2$. This is the shortest interesting profile in the sense that its high profile support 5 is larger than the number of symbols appearing. The proof of this follows from Theorem 45 on quasi-uniform profiles.

**Theorem 39**

$$\hat{P}_{2^11^2} = (1/5, 1/5, 1/5, 1/5, 1/5).$$

and for all patterns $\psi \in \Psi_{2^11^2}$

$$\hat{P}(\psi) = \frac{12}{125}.$$ \hfill \Box

### 15.2 Skewed profiles

Consider the skewed profile $\bar{\varphi} = r^11^u$, namely, one symbol appearing $r$ times and $u$ symbols appearing once each. We will show that for certain values of $r, u$, $\hat{P}_{\bar{\varphi}} = \left(\frac{r}{r+u}\right)$, a mixed distribution comprising a single symbol with probability $\frac{r}{r+u}$ and a continuous part with probability $\frac{u}{r+u}$. In the following two lemmas we derive conditions of $\hat{P} = (p_1, p_2, \ldots)$, the high profile distribution of the profile $\bar{\varphi} = r^11^u$. The conditions are an upper bound on $p_1$ when $p_2 = 0$ and a lower bound on $p_1$. For sufficiently large $u$, the upper bound can be made smaller than the lower bound leading to the conclusion that in such cases $p_2$ must be 0. If $p_2 = 0$ it is easy to show that $p_1 = \frac{r}{r+u}$.

**Lemma 40** If $\hat{P} = (p_1, p_2, \ldots) \in \mathcal{P}_m$ is the high-profile distribution of $\bar{\varphi} = r^11^u$, then $p_2 = 0$, or

$$p_1 \leq \frac{\sqrt{2u - 1} - 1}{u - 1}.$$

**Proof** We construct a distribution $Q = (p_1, p_3, \ldots)$ where the discrete symbol with probability $p_2$ is replaced by adding the probability $p_2$ to the continuous part. Consider $\psi = 1^r23\ldots u + 1$, the canonical pattern of $\bar{\varphi}$. We will show that $Q$ assigns a larger probability that $\hat{P}$ to $\psi$ unless $p_2 = 0$ or

$$p_1 \leq \frac{\sqrt{2u - 1} - 1}{u - 1}.$$ 

Now from (34)

$$\hat{P}(1^r23\ldots u + 1) = \sum_i p_i^r \hat{P}_{-i}(12\ldots u) \tag{48}$$

56
and

\[ Q(1^r2\ldots u + 1) = \sum_{i \neq 2} p_i^r Q^{-i}(12\ldots u). \]  \hfill (49)

Note that for all \( i \neq 2 \)

\[ \hat{P}_{-i}(12\ldots u) = \hat{P}_{-\{i,2\}}(12\ldots u) + up_2 \hat{P}_{-\{i,2\}}(12\ldots u - 1) \]  \hfill (50)

while, by construction,

\[ Q_{-i}(\psi)(12\ldots u) = \sum_{j=0}^u \binom{u}{j} p_j^r \hat{P}_{-\{i,2\}}(12\ldots u - j) \]

\[ \geq \hat{P}_{-\{i,2\}}(12\ldots u) + up_2 \hat{P}_{-\{i,2\}}(12\ldots u - 1) + \frac{u(u - 1)}{2} p_2^r \hat{P}_{-\{i,2\}}(12\ldots u - 2). \]  \hfill (51)

Combining (48–51) we obtain

\[ \hat{P}(1^r2\ldots u + 1) - Q(1^r2\ldots u + 1) \leq p_2^r \hat{P}_{-2}(12\ldots u) - \frac{u(u - 1)}{2} \sum_{i \neq 2} p_i^r p_2^r \hat{P}_{-\{i,2\}}(12\ldots u - 2) \]

\[ \leq \left( (1 - p_1)(1 + (u - 1)p_1)p_2^r - \frac{u(u - 1)}{2} p_1^r p_2^r \right) \hat{P}_{-\{1,2\}}(12\ldots u - 2) \]

\[ \leq \left( (1 - p_1)p_2^{r-2} - \frac{u - 1}{2} p_1^r \right) up_2^r \hat{P}_{-\{1,2\}}(12\ldots u - 2) \]  \hfill (52)

where the second inequality is obtained by omitting all the terms in the summation except the one corresponding to \( i = 1 \), and by noting that

\[ \hat{P}_{-2}(12\ldots u) = \hat{P}_{-\{1,2\}}(12\ldots u) + up_1 \hat{P}_{-\{1,2\}}(12\ldots u - 1) \]

\[ \leq (1 - p_1)^2 \hat{P}_{-\{1,2\}}(12\ldots u - 2) + up_1(1 - p_1) \hat{P}_{-\{1,2\}}(12\ldots u - 2) \]

\[ \leq (1 - p_1)(1 + (u - 1)p_1) \hat{P}_{-\{1,2\}}(12\ldots u - 2). \]

Since \( \hat{P} \) is the high-profile distribution,

\[ \hat{P}(1^r2\ldots u + 1) - Q(1^r2\ldots u + 1) \geq 0 \]

and therefore from (52) either \( p_2 = 0 \) or

\[ (1 - p_1)p_2^{r-2} - \frac{u - 1}{2} p_1^r \geq 0. \]

Since \( p_2 \leq p_1 \), this implies that

\[ (u - 1)p_1^2 + 2p_1 - 2 \leq 0, \]

\( i.e.\)

\[ p_1 \leq \frac{\sqrt{2u - 1} - 1}{u - 1}. \]  \hfill \( \Box \)

**Lemma 41** If \( \hat{P} = (p_1, p_2, \ldots) \in \mathcal{P}_M \) is the high-profile distribution of \( \hat{\varphi} = \varphi^1 \varphi^u \), then

\[ p_1 \geq \frac{r - 1}{r + u - 1} - \left( \frac{\ln 2}{2} \frac{1}{r + u - 1} \log \left( \frac{r + u}{r} \right) + D\left( \frac{r - 1}{r + u - 1} \parallel \frac{r}{r + u} \right) \right)^{1/2}. \]

57
**Proof** If \( p_1 > \frac{r-1}{r+u-1} \), the result holds trivially. Therefore we are left to consider the case when \( p_1 \leq \frac{r-1}{r+u-1} \). Consider the distribution \( P = \left( \frac{r}{r+u} \right) \), namely one discrete symbol with probability \( \frac{r}{r+u} \) and a continuous part with probability \( \frac{u}{r+u} \). Then
\[
P(1^r \ldots u + 1) = \left( \frac{r}{r+u} \right)^r \left( \frac{u}{r+u} \right)^u.
\]

Let
\[
1^r \tilde{u} \overset{\text{def}}{=} \{1^r \tilde{\psi}_3 \ldots \tilde{\psi}_{u+1} : \tilde{\psi}_i \neq 1, i \geq 2 \}
\]
denote the set of all patterns where 1 occurs \( r \) times followed by \( u \) integers, none of which equal 1. Clearly \( 1^r \tilde{u} \ldots u + 1 \) belongs to the set. Hence
\[
\left( \frac{r}{r+u} \right)^r \left( \frac{u}{r+u} \right)^u = P(1^r \tilde{u} \ldots u + 1) \leq \hat{P}(1^r \tilde{u} \ldots u + 1) \leq \hat{P}(1^r \tilde{u})
\]
\[
\leq \sum_i p_i (1-p_i)^u
\]
\[
\leq \sum_i p_i p_{i-1} (1-p_i)^u
\]
\[
\leq p r^{-1} (1-p)^u
\]
where \((a)\) follows from \( p_1 \leq \frac{r-1}{r+u-1} \). This can be rewritten as
\[
\frac{r}{r+u} 2^{-(r+u-1)D\left( \frac{r-1}{r+u-1} \mid \mid \frac{r}{r+u} \right)} \leq 2^{-(r+u-1)D\left( \frac{r-1}{r+u-1} \mid \mid \frac{r}{r+u} \right)}
\]
where \( h(p) = -p \log p - (1-p) \log (1-p) \) is the binary entropy function and
\[
D(p||q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}
\]
is the Kullback-Leibler divergence between two Bernoulli distributions with parameters \( p \) and \( q \). Using Pinsker’s inequality (see e.g., Lemma 12.6.1 in [?]), and rearranging the terms in (53) we obtain
\[
\left( \frac{r-1}{r+u-1} - p_1 \right)^2 \leq \frac{\ln 2}{2} D\left( \frac{r-1}{r+u-1} \mid \mid p_1 \right)
\]
\[
\leq \frac{\ln 2}{2} \left( \frac{1}{r+u-1} \log \left( \frac{r+u}{r} \right) + D\left( \frac{r-1}{r+u-1} \mid \mid r \right) \right).
\]
Hence the Lemma. \( \square \)

Combining these two Lemmas we show that the high-profile distribution of most skewed profiles is \( \left( \frac{r}{r+u} \right) \).

**Theorem 42** For all skewed profiles \( \varphi = r^11^u \) such that
\[
\frac{r-1}{r+u-1} - \left( \frac{\ln 2}{2} \left( \frac{1}{r+u-1} \log \left( \frac{r+u}{r} \right) + D\left( \frac{r-1}{r+u-1} \mid \mid r \right) \right) \right)^{1/2} > \sqrt{2u-1 - 1 \over u - 1}
\]
the high-profile distribution is \( \hat{P} = \left( \frac{r}{r+u} \right) \).

**Proof** It is easy to verify that of all the distributions with \( p_2 = 0 \) the one that assigns the largest probability to \( \varphi = r^11^u \) is \( \left( \frac{r}{r+u} \right) \). From Lemmas 40 and 41, if
\[
\frac{r-1}{r+u-1} - \left( \frac{\ln 2}{2} \left( \frac{1}{r+u-1} \log \left( \frac{r+u}{r} \right) + D\left( \frac{r-1}{r+u-1} \mid \mid r \right) \right) \right)^{1/2} > \sqrt{2u-1 - 1 \over u - 1}
\]
then \( p_2 = 0 \). \( \square \)
Corollary 43  For all sufficiently large $u$, if $r \geq 2\sqrt{u}$ then $\hat{P}_{r,1} = \left( \frac{r}{r+u} \right)$.

15.3 Quasi-uniform profiles

Uniform profiles are of the form $r^m$, i.e., each of the $m$ symbols appears $r$ times. Generalizing this notion slightly, quasi-uniform profiles are profiles of the form $\bar{r} = (r+1)^{m-t}$. In Theorem 45, we show that the high-profile distribution of all quasi-uniform profiles except the all-distinct profile is a uniform discrete distribution. While this may not be surprising when $r \geq 2$, the proof is non-trivial. The case $r = 1$ is particularly interesting. Based on Theorem 42, and its proof one may be tempted to guess that the high-profile distribution of patterns such as 1123, 11234 are mixed. However, as we show in Theorem 45, the high-profile distributions are discrete.

To prove the result we require Lemma 44 which concerns discrete high profile distributions with finite support size. Let $\mathcal{P}_F \subset \mathcal{P}_M$ denote the set of discrete distributions with finite support.

**Lemma 44**  For all patterns $\bar{\psi}$, if $\hat{P}$ achieves $\sup_{P \in \mathcal{P}_F} P(\bar{\psi})$ then $\hat{P}$ is a high-profile distribution of $\bar{\psi}$.

**Proof**  Let $\hat{P}$ be a high-profile distribution of $\bar{\psi}$. Then

$$\hat{P}(\bar{\psi}) = \sup_{P \in \mathcal{P}_M} P(\bar{\psi}).$$

From Corollary 24, $\hat{P} \in \mathcal{P}_R$ or $\hat{P}$ is mixed with finite discrete size $\hat{k}$. In the former case, the lemma follows trivially. In the latter case, we will show that for all $\epsilon > 0$

$$\hat{P}(\bar{\psi}) > \hat{P}(\bar{\psi}) - \epsilon$$

and this proves the Lemma.

Let $\bar{\psi} = 1^{\mu_1}2^{\mu_2} \ldots m^{\mu_m}$ be canonical of length $n$. Let $\hat{P} = (p_1, p_2, \ldots, p_k)$ and let $q$ be its continuous probability. Since $q > 0$, from Theorem 21, the prevalence of 1, $\varphi_1$, in $\bar{\psi}$ must be strictly greater than 1. For $0 \leq i \leq \varphi_1$, let $\bar{\psi}_{-i}$ denote the pattern $1^{\mu_1}2^{\mu_2} \ldots (m-i)^{\mu_{m-i}}$. Let

$$\hat{P}_d = \left( \frac{p_1}{1-q}, \frac{p_2}{1-q}, \ldots, \frac{p_k}{1-q} \right)$$

denote the distribution consisting of the discrete part of $\hat{P}$ normalized to sum to 1. Since $\bar{\psi}$ is canonical

$$\hat{P}(\bar{\psi}) = (1-q)^n \hat{P}_d(\bar{\psi}) + \sum_{i=1}^{\varphi_1} \left( \frac{\varphi_1}{i} \right) q^i (1-q)^{n-i} \hat{P}_d(\bar{\psi}_{-i}).$$

(54)

Consider the uniform distribution $\hat{P}_\mu$ over a set of $\left\lceil \frac{n^2 P(\bar{\psi})}{2\epsilon} \right\rceil$ symbols that is disjoint from the support of $\hat{P}_d$. We construct the discrete mixture distribution $Q$ where one of $\hat{P}_d$ and $\hat{P}_\mu$ is selected with probabilities $1-q$ and $q$ respectively. If this procedure is repeated $n$ times, to draw the i.i.d. sequence $\bar{X} = X_1X_2 \ldots X_n$, then observe that

$$Q(\bar{X} = \bar{\psi}) > (1-q)^n \hat{P}_d(\bar{\psi}) + \sum_{i=1}^{\varphi_1} \left( \frac{\varphi_1}{i} \right) q^i (1-q)^{n-i} \hat{P}_d(\bar{\psi}_{-i}) \hat{P}_\mu(12 \ldots i)$$

(55)

59
where the \(i\)th term on the right, \(0 \leq i \leq \varphi_1\), corresponds to the probability of selecting \(\hat{P}_d\) the first \(n - i\) times, \(\hat{P}_u\) the last \(i\) times, and observing the pattern \(\tilde{\psi}\). Now \(\hat{P}_u(12\ldots i)\) is the probability of drawing all distinct symbols in \(i\) attempts when \(\hat{P}_u\) is the underlying distribution. The probability that a given symbol in \(\hat{P}_u\)'s support appears more than once when \(i\) symbols are drawn according to \(\hat{P}_u\) is at most \(\frac{i(i-1)}{2} \left[ \frac{n^2\hat{P}(\tilde{\psi})}{2e} \right]^2\). Therefore for \(i = 1\), \(\hat{P}_u(12\ldots i) = 1\) and for \(i \geq 2\),

\[
\hat{P}_u(12\ldots i) \geq 1 - \frac{n^2\hat{P}(\tilde{\psi})}{2e} \frac{i(i-1)}{2} \left[ \frac{n^2\hat{P}(\tilde{\psi})}{2e} \right]^2 \geq 1 - \frac{\epsilon}{\hat{P}(\tilde{\psi})}.
\]

Substituting this in (55) and comparing with (54) we obtain

\[
Q(\tilde{\psi}) > (1-q)^n \hat{P}_d(\tilde{\psi}) + \sum_{i=1}^{\varphi_1} \left( \varphi_1 \right)_i q^i (1-q)^{n-i} \hat{P}_d(\tilde{\psi}^{-i}) \left( 1 - \frac{\epsilon}{\hat{P}(\tilde{\psi})} \right)
\]

\[
> (1-q)^n \hat{P}_d(\tilde{\psi}) + \sum_{i=1}^{\varphi_1} \left( \varphi_1 \right)_i q^i (1-q)^{n-i} \hat{P}_d(\tilde{\psi}^{-i}) \left( 1 - \frac{\epsilon}{\hat{P}(\tilde{\psi})} \right)
\]

\[
= \hat{P}(\tilde{\psi}) \left( 1 - \frac{\epsilon}{\hat{P}(\tilde{\psi})} \right) = \hat{P}(\tilde{\psi}) - \epsilon.
\]

Clearly \(Q \in \mathcal{P}_r\) and therefore

\[
\hat{P}(\tilde{\psi}) \geq Q(\tilde{\psi}) > \hat{P}(\tilde{\psi}) - \epsilon.
\]

Before we proceed to prove our main result on quasi-uniform profiles, we require the following notation. Extending the definition in Section 14.2 for any pattern \(\tilde{\psi} = 1^{\mu_1}2^{\mu_2} \ldots m^{\mu_m}\) and integers \(1 \leq \ell, r \leq m\), define the pattern \(\tilde{\psi}_{-(\ell, r)}\) to be \(1^{\mu_1} \ldots (\ell - 1)^{\mu_{\ell-1}} (\ell^r+1) \ldots (r-1)^{\mu_{r-1}} (r^m+1) \ldots (m-2)^{\mu_m}\), the pattern obtained by removing \(\ell\) and \(r\) from \(\tilde{\psi}\). For example, if \(\tilde{\psi} = 1^{\mu_1}2^{\mu_2}3^{\mu_3}4^{\mu_4}\) then \(\tilde{\psi}_{-2} = 1^{\mu_1}2^{\mu_2}3^{\mu_3}\) and \(\tilde{\psi}_{-(2, 3)} = 1^{\mu_1}2^{\mu_4}\). Then for all \(i < j\), and any pattern \(\tilde{\psi}\) with \(m\) symbols

\[
P(\tilde{\psi}) = P_{-(i,j)}(\tilde{\psi}) + \sum_{\ell=1}^{m} (p_1^{\mu_1} + p_2^{\mu_2}) P_{-(i,j)}(\tilde{\psi}_{-\ell}) + \sum_{\ell=1}^{m-1} \sum_{r>\ell} (p_1^{\mu_1} p_2^{\mu_2} + p_2^{\mu_2} p_1^{\mu_1}) P_{-(i,j)}(\tilde{\psi}_{-(\ell, r)}).
\]

(56)

**Theorem 45**  For all \(t \geq 1\) and \(r \geq 1\)

\[
\hat{P}_{(r+1)r^{m-1}} = \left( \frac{1}{k}, \ldots, \frac{1}{k} \right)
\]

where

\[
k = \min \left\{ k \geq m : \left( 1 + \frac{1}{k} \right)^{r^{m+t}} \left( 1 - \frac{m}{k+1} \right) > 1 \right\}.
\]

(57)

**Proof**  Without loss of generality let \(\tilde{\psi} = 1^{r+1}\ldots r^{r+1}(t+1)^{r}\ldots m^{r}\). Recall that \(\mathcal{P}_r\) is the set of all discrete distributions with finite support. We will show that the supremum of \(P(\tilde{\psi})\) over \(\mathcal{P}_r\) is achieved by a uniform distribution over \(k\) where \(k\) is given by (57) and the theorem will follow from Lemma 44.

Consider \(P = (p_1, \ldots, p_k)\) that maximizes the probability of the pattern \(\tilde{\psi}\) over all distributions with support size \(k\) with the property that for all \(1 \leq i \leq k\), \(p_i > 0\). If the support size
of the distribution that maximizes the probability of $\tilde{\psi}$ is less than $k$, decrease $k$ until such a distribution is found. Then by Kuhn-Tucker conditions, for all $1 \leq i, j \leq k$

$$g(P) \overset{\text{def}}{=} \frac{\partial P(\tilde{\psi})}{\partial p_i} - \frac{\partial P(\tilde{\psi})}{\partial p_j} = 0.$$  

Note that for $0 < t < m$, $\tilde{\psi}_1 = 1^{r+1} \ldots (t-1)r^{t-1}r \ldots (m-1)r$, $\tilde{\psi}_m = 1^{r+1} \ldots t^{r+1}(t+1)^r \ldots (m-1)^r$, $\tilde{\psi}_{-1, m} = 1^{r+1} \ldots (t-1)r^{t-1}r \ldots (m-2)r$, $\tilde{\psi}_{-1, 2} = 1^{r+1} \ldots (t-1)r^{t+1}(t-2)^r \ldots (m-2)^r$. Hence

$$\frac{\partial P(\tilde{\psi})}{\partial p_i} = (m-t)r^{t-1}r p_i^{r-1}P_{-(i,j)}(\tilde{\psi}_m) + (m-t)(m-t-1)r^{t-1}r p_i^{r-1}p_j^{r-1}P_{-(i,j)}(\tilde{\psi}_{m-1, m})$$

$$+ t(r+1)P_{-(i,j)}(\tilde{\psi}_1) + t(t-1)(r+1)p_i^{r+1}P_{-(i,j)}(\tilde{\psi}_{1,2})$$

$$+ t(m-t)(p_i^{r-1}p_j^{r+1} + (r+1)p_i^{r+1}p_j^{r+1})P_{-(i,j)}(\tilde{\psi}_{1,1}).$$

Hence

$$g(P) = (m-t)r(p_i^{r-1} - p_j^{r-1})P_{-(i,j)}(\tilde{\psi}_m)$$

$$+ (m-t)(m-t-1)r(p_i^{r-1}p_j^{r-1} - p_j^{r-1}p_i^{r-1})P_{-(i,j)}(\tilde{\psi}_{m-1, m})$$

$$+ t(r+1)(p_i^{r-1}p_j^{r-1})P_{-(i,j)}(\tilde{\psi}_1) + t(t-1)(r+1)(p_i^{r+1}p_j^{r+1} - p_j^{r+1}p_i^{r+1})P_{-(i,j)}(\tilde{\psi}_{1,2})$$

$$+ t(m-t)(p_i^{r-1}p_j^{r+1} - p_j^{r-1}p_i^{r+1})P_{-(i,j)}(\tilde{\psi}_{1,1}).$$

Consider a new distribution $Q$ of smaller support size obtained by adding $p_i$ and $p_j$, i.e., $Q$ is comprised of the discrete probabilities $p_1, \ldots, p_{i-1}, p_i, p_{i+1}, \ldots, p_{j-1}, p_j, \ldots, p_k$, and $p_i + p_j$ suitably sorted. Then

$$P(\tilde{\psi}) - Q(\tilde{\psi}) = t(t-1)(p_ip_j)^{r+1}P_{-(i,j)}(\tilde{\psi}_{1,2}) + (m-t)(m-t-1)(p_ip_j)^{r+1}P_{-(i,j)}(\tilde{\psi}_{m-1, m})$$

$$+ t(m-t)(p_i^{r+1}p_j^{r+1} + p_j^{r+1}p_i^{r+1})P_{-(i,j)}(\tilde{\psi}_{1,1})$$

$$- (m-t) \sum_{\ell=1}^{r-1} \binom{r}{\ell} p_i^{r-\ell}p_j^{\ell}P_{-(i,j)}(\tilde{\psi}_m) - t \sum_{\ell=1}^{r} \binom{r+1}{\ell} p_i^{r+1-\ell}p_j^{\ell}P_{-(i,j)}(\tilde{\psi}_1).$$

Comparing (58) and (59) and noting that when $i > j$, $p_i \leq p_j$, we obtain

$$p_ip_j g(P) \geq r(p_j - p_i)(P(\tilde{\psi}) - Q(\tilde{\psi}))$$

with equality iff $t = 1$ and $r = 1$. Therefore if $g(P) = 0$ then $p_i = p_j$ for all $i, j$, or there exists another distribution with smaller support size that assigns a greater probability to $\tilde{\psi}$. Therefore if $\mathcal{P}_U \subset \mathcal{P}_F$ denotes the set of all discrete uniform distributions then

$$\sup_{P \in \mathcal{P}_U} P(\tilde{\psi}) = \sup_{P \in \mathcal{P}_U} P(\tilde{\psi}).$$

(In fact, if $t > 1$ or $r > 1$ and there exists a distribution $P \in \mathcal{P}_F$ that achieves the supremum on the left hand side then $P \in \mathcal{P}_U$.) Further consider a uniform distribution $P$ of support size $k$. Then

$$P(\tilde{\psi}) \leq P(11) = \frac{1}{k}.$$  

However the probability assigned by a uniform distribution over $m$ to $\tilde{\psi}$ is

$$\frac{m!}{m^m + t}.$$
which implies that
\[
\sup_{P \in \mathcal{P}_U} P(\bar{\psi}) = \sup_{P \in \mathcal{P}_U : t \leq \frac{m r + t}{m}} P(\bar{\psi}),
\]
namely the supremum over all uniform distributions whose support size is bounded by a function of \(m, r,\) and \(t\). Since this supremum is clearly achieved by some uniform distribution it follows from Lemma 44 that the high-profile distribution of \(\bar{\psi}\) is discrete with finite support size. Let \(P_{u,k}\) denote the uniform distribution with support size \(k\). Then
\[
P_{u,k}(\bar{\psi}) = \frac{k m}{k r + t}.
\]
It is easy to verify that this expression either decreases with \(k\) or increases with \(k\) and then decreases. Therefore the maximum is attained at the smallest \(k\) such that
\[
\frac{P_{u,k}(\bar{\psi})}{P_{u,k+1}(\bar{\psi})} > 1,
\]
which is given by (57).

In the following corollary we derive \(\hat{t}\) for quasi-uniform profiles \((r + 1)^{t r - t}\) when \(m\) tends to infinity.

**Corollary 46** For the sequence of profiles \(\{(r + 1)^{t(m) r - t(m)}\}\) where
\[
\lim_{m \to \infty} m^{-1} t(m) = \beta,
\]
if \(r + \beta > 1\)
\[
\lim_{m \to \infty} \frac{t(r+1)^{t(m) r - t(m)}}{m} = \alpha
\]
where
\[
-\alpha \ln \left(1 - \frac{1}{\alpha}\right) = r + \beta,
\]
and if \(t(m) = t \geq 1\) is constant and \(r = 1\)
\[
\lim_{m \to \infty} \frac{t(r+1)^{t(m) r - t(m)}}{m^2} = \frac{1}{2t}.
\]

**Proof** Let \(\hat{t} = \hat{\alpha} m\). From Theorem 45,
\[
\hat{\alpha} = \min \left\{ \alpha \geq 1 : \alpha m \in \mathbb{P}, \left(1 + \frac{1}{\alpha m}\right)^{r m + t(m)} \left(1 - \frac{m}{\alpha m + 1}\right) > 1 \right\}.
\]
For sufficiently large \(m\) we are seeking \(\alpha\) such that
\[
\left(1 + \frac{1}{\alpha m}\right)^{r m + t(m)} \left(1 - \frac{m}{\alpha m + 1}\right) = 1.
\]
Taking logarithms on both sides and letting \(m\) tend to infinity we obtain
\[
\lim_{m \to \infty} \left( r + \frac{t(m)}{m} \right) \ln \left(1 + \frac{1}{\alpha m}\right)^m = -\lim_{m \to \infty} \ln \left(1 - \frac{m}{\alpha m + 1}\right).
\]
This reduces to
\[
r + \beta = -\alpha \ln \left(1 - \frac{1}{\alpha}\right),
\]
62
which has a unique solution when \( r + \beta > 1 \).

We consider the case when \( r = 1 \) and \( t(m) = t \geq 1 \) is constant. Observe that as \( m, k \) tend to infinity
\[
(1 + \frac{1}{k})^{m+t} \left(1 - \frac{m}{k+1}\right) = 1 + \frac{t}{k} - \frac{m^2}{2k^2} + \frac{m}{2k} - \frac{m^3}{3k^3} + O\left(\frac{1}{k^2}\right) + O\left(\frac{m^2}{k^3}\right).
\]
For sufficiently large \( m \) if
\[
m \leq k < \frac{m^2}{2t}
\]
this quantity is strictly less than 1, and if \( k = \frac{m^2}{2t} (1 + o(1)) \) the quantity is greater than 1. This proves the second part of the Corollary.

This result also leads to an interesting observation and conjecture about the largest possible discrete size of a high-profile distribution.

**Corollary 47** There exists a sequence of profiles \( \{\bar{\psi}_m\} \) with \( m \) symbols for which
\[
\lim_{m \to \infty} \frac{k\bar{\psi}_m}{m^2} = \frac{1}{2}.
\]
The sequence of profiles that satisfy the above corollary are \( \bar{\psi}_m = 2^{1 \cdot m^{-1}} \). Note that this profile is also a skewed profile and therefore one might be tempted to guess, as in Corollary 43, that a mixed distribution with one discrete symbol might maximize the pattern probability. Consider the pattern \( \bar{\psi} = 1^2 23 \ldots m \) whose profile is \( \bar{\psi}_m = 2^{1 \cdot m^{-1}} \). It is easy to see that the mixed distribution with one discrete symbol that maximizes the probability of this pattern is \( P_m = \left(\frac{2}{m+1}\right) \) and the maximum probability is
\[
P_m(\bar{\psi}) = \left(\frac{2}{m+1}\right)^2 \left(\frac{m-1}{m+1}\right)^{m-1} = \left(\frac{2}{m+1}\right)^2 \left(1 - \frac{2}{m+1}\right)^{m-1}.
\]
Note that for all \( x > -1 \),
\[
\log(1 + x) \leq x.
\]
Therefore for all \( m > 3 \)
\[
P_m(\bar{\psi}) \leq \left(\frac{2}{m+1}\right)^2 e^{-\frac{2(m-1)}{m+1}} \leq \left(\frac{2}{e(m+1)}\right)^2 e^{\frac{4}{m+1}}.
\]
On the other hand, the probability assigned to \( \bar{\psi} \) by \( P_u \), a uniform distribution over \( \frac{m^2}{2} \) symbols is
\[
P_u(\bar{\psi}) = \frac{\left(\frac{m^2}{2}\right)!}{\left(\frac{m^2}{2} - m\right)! \left(\frac{m^2}{2}\right)^{m+1}}.
\]
Stirling’s approximation gives us that for all \( n \)
\[
\sqrt{2\pi n} \left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}.
\]
Applying this we obtain that
\[
P_u(\bar{\psi}) \geq e^{-m} \sqrt{\frac{m^2 - m}{m^2 - 2m}} \left(\frac{m^2}{m^2 - 2m}\right)^{\frac{m^2 - m}{m^2} \frac{2}{m^2} e^{-\frac{1}{6m^2 - 12m}}}.
\]
Observing that for all $x > 0$
\[ \log(1 + x) > x - \frac{x^2}{2} \]
we obtain that for all $m > 2$
\[
P_u(\psi) > e^{-m} \left( 1 + \frac{2m}{m^2 - 2m} \right) \frac{m^2}{m^2} \frac{2}{m^2} e^{-\frac{1}{6m^2 - 12m}}
\]
\[
> e^{-m} \left( \frac{\frac{1}{m - 1} - \frac{1}{m}}{\frac{1}{m} - 1} \right) \frac{\frac{2}{m^2}}{m^2} e^{-\frac{1}{6m^2 - 12m}}
\]
\[
= \frac{2}{e^m} e^{-\left( \frac{2}{m^2 + \frac{1}{6m^2 - 12m}} \right)}.
\]
Comparing this with (60), we obtain that for all $m > 3$,
\[
P_u(\psi) > P_m(\psi)
\]
when
\[
\frac{2}{e} < e^{-\left( \frac{2}{m^2 + \frac{1}{6m^2 - 12m}} + \frac{1}{m+1} \right)}.
\]
It can be verified that this condition holds for all $m \geq 20$.

In general, it can be expected that if the number of symbols appearing more than once is increased when keeping the profile length constant, the discrete size of $\hat{P}$ would decrease. The only profile with fewer symbols appearing more than once than $\bar{\varphi}_m = 2^1 1^{m-1}$ is the all-distinct profile, whose $\hat{P}$ is continuous and therefore $\hat{k} = 0$. This leads to the following conjecture.

For any sequence of profiles $\{\bar{\varphi}_m\}$ with $m$ symbols
\[
\lim_{m \to \infty} \frac{k_{\bar{\varphi}_m}}{m^2} \leq \frac{1}{2}.
\]

15.4 Binary Profiles

In this section we determine the high-profile distribution of a simple class of profiles, those with only two symbols. To do so, we use the following result by Alon [?] on the maximization of a 2-variable symmetric polynomial.

**Lemma 48** [?] For all $n \geq 2$ and $1 \leq n_0 \leq n/2$, if $((n - n_0) - n_0)^2 \leq n$, then for $0 \leq p \leq \frac{1}{2}$,
\[ f(p) \overset{\text{def}}{=} p^{n_0}(1-p)^{n-n_0} + (1-p)^{n_0}p^{n-n_0} \] (61)
is maximized by $p = 1/2$, otherwise the polynomial
\[ n_0 \cdot x^{n-2n_0+1} - (n - n_0) \cdot x^{n-2n_0} + (n - n_0) \cdot x - n_0 \]
has a unique root $\alpha$ in $(0,1)$ and $f(p)$ is maximized by $p = 1/\alpha$.  

**Proof** The proof, not given in [?], is provided for completeness. Differentiating $f(p)$ we obtain
\[
f'(p) = n_0 p^{n_0-1}(1-p)^{n-n_0} - (n - n_0)p^{n_0}(1-p)^{n-n_0-1}
\]
\[
+ (n - n_0)p^{n-n_0-1}(1-p)^{n_0} - n_0p^{n-n_0}(1-p)^{n_0-1}.
\]
The above equation can be rewritten as
\[ g(\alpha) \overset{\text{def}}{=} \left( \frac{1}{1 + \alpha} \right)^{n-1} \left( n_0 \alpha^{n_0-1} - (n - n_0)\alpha^{n_0} + (n - n_0)\alpha^{n-n_0-1} - n_0\alpha^{n-n_0} \right), \]
where \( p = \frac{\alpha}{1+\alpha} \). Since \( p \in [0,1/2] \), \( \alpha \in [0,1] \). If \( g(\alpha) = 0 \), \( \alpha \) is a root of

\[
 n_0 \cdot x^{n-2n_0 + 1} - (n - n_0) \cdot x^{n-2n_0} + (n - n_0) \cdot x - n_0.
\] (62)

We first deal with the trivial cases where, \( n - 2n_0 = 0 \) and \( n - 2n_0 = 1 \). If \( n - 2n_0 = 0 \), then this reduces to

\[ n(x - 1), \]

and if \( n - 2n_0 = 1 \) the polynomial reduces to

\[ n_0(x^2 - 1). \]

In both cases the only positive root is \( x = 1 \), implying that \( f(p) \) is maximized at \( p = 1/2 \).

If \( n - 2n_0 > 1 \), then by Descartes’ rule of signs, the polynomial in (62) has at most 3 positive roots. Observe that 1 is always a root and if \( x \) is a root so is \( 1/x \). Therefore if there exists a root in \((0, 1)\) it has to be unique. We now show that if \(((n - n_0) - n_0)^2 < n \) then the function \( f \) is maximized when \( \alpha = 1 \). From the Taylor series of \( g(\alpha) \), it is easy to observe that for all sufficiently small \( \alpha \), \( g(\alpha) > 0 \).

The first derivative of \( g \) at 1 is given by,

\[
g'(1) = n_0(n_0 - 1) - (n - n_0)n_0 + (n - n_0)(n_0 - 1) - n_0(n - n_0) = ((n - n_0) - n_0)^2 - n.
\]

If \(((n - n_0) - n_0)^2 < n \), \( g'(1) < 0 \) and, since \( g(1) \) is always zero, \( g(1 - \alpha) > 0 \) for all sufficiently small \( \alpha \). Since, for sufficiently small \( \alpha \), both \( g(\alpha) \) and \( g(1 - \alpha) \) are \( > 0 \), it follows that \( g \) crosses the abscissa an even number of times between 0 and 1. However \( g(\alpha) \) has at most one root in \((0, 1)\) and therefore \( g \) cannot cross the abscissa at this root implying that it is not a maxima. Also, since \( g(1 - \alpha) < 0 \) and \( g(1 + \alpha) > 0 \) for sufficiently small \( \alpha \), \( \alpha = 1 \) corresponds to a local maxima of \( f \left( \frac{\alpha}{1+\alpha} \right) \). Therefore, \( f \) is maximized when \( \alpha = 1 \), i.e., \( p = 1/2 \).

If \(((n - n_0) - n_0)^2 = n \) then \( g'(1) = 0 \) and it can be verified that \( g''(1) = 0 \) and \( g'''(1) < 0 \) if 
\( n - n_0 \geq 3 \). Therefore there exists no root in \((0, 1)\), and \( \alpha = 1 \), i.e., \( p = 1/2 \) is a local maxima for \( f \). It can be verified that if \( n - n_0 \leq 2 \), then for all \( n \), \( n_0 \leq n/2 \), \( ((n - n_0) - n_0)^2 \neq n \).

However if \(((n - n_0) - n_0)^2 > n \) then \( g'(1) > 0 \) and therefore \( g(1 - \alpha) \) is \( < 0 \) for all sufficiently small \( \alpha \). Since \( g(\alpha) > 0 \) for all sufficiently small \( \alpha \) it follows that \( g(\alpha) \) crosses the abscissa an odd number of times in \((0, 1)\). Since by earlier arguments there can be at most one root between 0 and 1, it follows that this root is unique and corresponds to \( g(\alpha) \) changing signs from positive to negative. Therefore this root corresponds to a maxima of \( f \left( \frac{\alpha}{1+\alpha} \right) \). Also, since \( g(1 - \alpha) < 0 \) and \( g(1 + \alpha) > 0 \) for sufficiently small \( \alpha \), \( \alpha = 1 \) corresponds to a local minima of \( f(p) \). Therefore, \( f(p) \) is maximized at \( p = \frac{\alpha}{1+\alpha} \) where \( \alpha \) is the unique root of \( g(\alpha) \) in \((0, 1)\).

Using the Lemma we derive \( \hat{P} \) for all patterns containing exactly two symbols.

**Theorem 49** For all \( 1 \leq n_0 \leq n/2 \),

\[
\hat{P}_{(n-n_0)\cdot n_0} = \begin{cases} (0.5, 0.5) & ((n - n_0) - n_0)^2 \leq n, \\ \left( \frac{1}{1+\alpha}, \frac{\alpha}{1+\alpha} \right) & ((n - n_0) - n_0)^2 > n, \end{cases}
\]

and for all patterns of that profile

\[
\hat{P}(\vec{v}) = \begin{cases} \frac{1}{\alpha^{n_0 + n - n_0}} & ((n - n_0) - n_0)^2 \leq n, \\ \left( \frac{\alpha}{1+\alpha} \right)^n & ((n - n_0) - n_0)^2 > n, \end{cases}
\]

65
where $\alpha$ is the unique root in $(0,1)$ of the polynomial

$$n_0 \cdot x^{n-2n_0+1} - (n - n_0) \cdot x^{n-2n_0} + (n - n_0) \cdot x - n_0.$$

**Proof**  If $n_0 = 1$, then from Theorem 35, $\hat{t} = 2$. Otherwise $n - n_0 \geq n_0 \geq 2$ and $\log (m + 1) = \log 3 < 2$. Then, by Corollary 27, $\hat{t} = 2$. Therefore for all patterns of profile $n_0^1(n - n_0)^1$, $\hat{P}$ is of the form $P = (1 - p, p)$. For any pattern $\bar{\psi}$ with profile $(n - n_0)^1n_0^1$, $P(\bar{\psi}) = p^{n_0}(1-p)^{n-n_0} + p^{n-n_0}(1-p)^{n_0}$. Therefore $\hat{P} = (1 - p^*, p^*)$ where $p^*$ maximizes $P(\bar{\psi})$. The theorem follows from Lemma 48.

**Acknowledgements**

We thank Warwick DeLauney and Doug Jungreis for simplifying our original proof for the high-profile distribution of the pattern 112, Shengjun (Alan) Pan for help with the experiments, and Ian Abramson, Noga Alon, Ron Graham, Nikola Jevtić, and Sajama for helpful discussions.