The Maximum Likelihood Probability of Unique-Singleton, Ternary, and Length-7 Patterns

Jayadev Acharya
ECE Department, UCSD
Email: jayadev@ucsd.edu

Alon Orlitsky
ECE & CSE Departments, UCSD
Email: alon@ucsd.edu

Shengjun Pan
CSE Department, UCSD
Email: s1pan@ucsd.edu

***Abstract*—We derive several pattern maximum likelihood (PML) results, among them showing that if a pattern has only one symbol appearing once, its PML support size is at most twice the number of distinct symbols, and that if the pattern is ternary with at most one symbol appearing once, its PML support size is three. We apply these results to extend the set of patterns whose PML distribution is known to all ternary patterns, and to all but one pattern of length up to seven.**

I. Introduction

Estimating the distribution underlying an observed data sample has important applications in a wide range of fields, including statistics, genetics, system design, and compression.

Many of these applications do not require knowing the probability of each element, but just the collection, or *multiset* of probabilities. For example, in evaluating the probability that when a coin is flipped twice both sides will be observed, we don't need to know $p(heads)$ and $p(tails)$, but only the multiset $\{p(heads), p(tails)\}$. Similarly to determine the probability that a collection of resources can satisfy certain requests, we don't need to know the probability of requesting the individual resources, just the multiset of these probabilities, regardless of their association with the individual resources. The same holds whenever just the data "statistics" matters.

One of the simplest solutions for estimating this probability multiset uses *standard maximum likelihood (SML)* to find the distribution maximizing the sample probability, and then ignores the association between the symbols and their probabilities. For example, upon observing the symbols $@\wedge@$, SML would estimate their probabilities as $p(@) = 2/3$ and $p(\wedge) = 1/3$, and disassociating symbols from their probabilities, would postulate the probability multiset $\{2/3, 1/3\}$.

SML works well when the number of samples is large relative to the underlying support size. But it falls short when the sample size is relatively small. For example, upon observing a sample of 100 distinct symbols, SML would estimate a uniform multiset over 100 elements. Clearly a distribution over a large, possibly infinite number of elements, would better explain the data. In general, SML errs in never estimating a support size larger than the number of elements observed, and tends to underestimate probabilities of infrequent symbols.

Several methods have been suggested to overcome these problems. One line of work began by Fisher [1], and was followed by Good and Toulmin [2], and Efron and Thisted [3]. Bunge and Fitzpatric [4] provide a comprehensive survey of many of these techniques.

A related problem, not considered in this paper estimates the probability of individual symbols for small sample sizes. This problem was considered by Laplace [5], Good and Turing [6], and more recently by McAllester and Schapire [7], Shamir [8], Gemelos and Weissman [9], Jedynak and Khudanpur [10], and Wagner, Viswanath, and Kulkarni [11].

A recent information-theoretically motivated method for the multiset estimation problem was pursued in [12], [13], [14]. It is based on the observation that since we do not care about the association between the elements and their probabilities, we can replace the elements by their order of appearance, called the observation's *pattern*. For example the pattern of $@\wedge@$ is 121, and the pattern of *abracadabra* is 12314151231.

Slightly modifying SML, this *pattern maximum likelihood (PML)* method asks for the distribution multiset that maximizes the probability of the observed pattern. For example, the 100 distinct-symbol sample above has pattern 123...100, and this pattern probability is maximized by a distribution over a large, possibly infinite support set, as we would expect. And the probability of the pattern 121 is maximized, to $1/4$, by a uniform distribution over two symbols, hence the PML distribution of the pattern 121 is the multiset $\{1/2, 1/2\}$.

To evaluate the accuracy of PML we conducted the following experiment. We took a uniform distribution over 500 elements, shown in Figure 1 as the solid (blue) line. We sampled the distribution with replacement 1000 times. In a typical run, of the 500 distribution elements, 6 elements appeared 7 times, 2 appeared 6 times, and so on, and 77 did not appear at all as shown in the figure. The standard ML estimate, which always agrees with empirical frequency, is shown by the dotted (red) line. It underestimates the distribution's support size by over 77 elements and misses the distribution's uniformity. By contrast, the PML distribution, as approximated by the EM algorithm described in [14] and shown by the dashed (green) line, performs significantly better and postulates essentially the correct distribution.

As shown in the above and other experiments, PML's empirical performance seems promising. In addition, several results have proved its convergence to the underlying distribution [13], yet analytical calculation of the PML distribution for specific patterns appears difficult. So far the PML distribution has been derived for only very simple or short patterns.

Among the simplest patterns are the *binary* patterns, consisting of just two distinct symbols, for example 11212. A formula for the PML distributions of all binary patterns was

978-1-4244-4313-0/09/$25.00 ©2009 IEEE

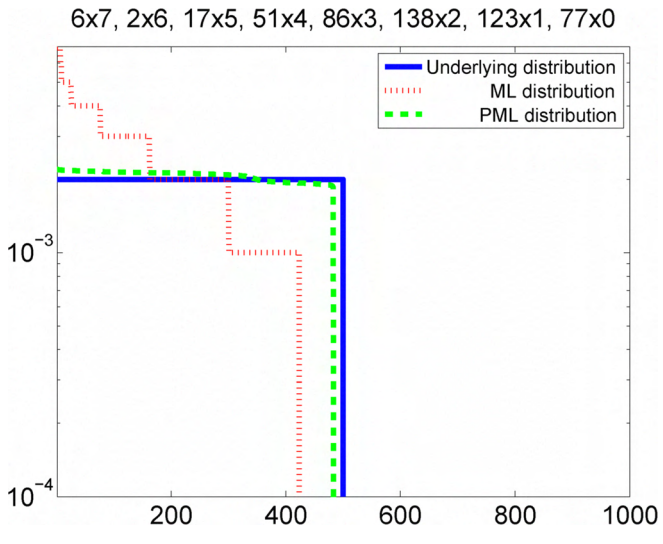

Fig. 1. SML and PML reconstruction of uniform distribution over 500 symbols from 1000 samples

derived in [12].

Another simple group of patterns are the *ternary* patterns, consisting of three distinct symbols, for example 121232. The PML distribution of some ternary patterns follows from results proven earlier. But so far not all ternary patterns have known PML's. In this paper we determine the PML of all previously unknown ternary patterns.

One of the most interesting applications of PML is to determine the underlying distribution's support size. The support size is of interest in many applications and is useful in simulations. Several bounds on the support size have been proven in [12]. We extend known bounds to show that if only one symbol in the sample appears once, then the PML support size is at most twice the number of distinct symbols.

We can apply the results described above to establish the PML distribution of many simple patterns, in particular we extend the set of patterns with known PML distributions to all but one pattern of length at most seven.

II. Notation

The *pattern* $\psi(\overline{x})$ of a sequence $\overline{x} \stackrel{\text{def}}{=} x_1^n$ is the integer sequence obtained by replacing each symbol x in $\overline{x}$ by the number of distinct symbols up to (and including) x's first appearance. For example, $\psi(abracadabra) = 12314151231$.

We denote the length of a pattern by n and its number of distinct symbols by m. The *multiplicity* of an integer ψ in a pattern $\overline{\psi}$ is the number μ_ψ of times ψ appears in $\overline{\psi}$. For example, for 12314151231, $n = 11$, $m = 5$, $\mu_1 = 5$, $\mu_2 = \mu_3 = 2$, and $\mu_4 = \mu_5 = 1$.

For simplicity, if a number ψ repeats consecutively i times, we abbreviate it as ψ^i. For example, we may write the pattern 11222111 as $1^2 2^3 1^3$. A pattern of the form $1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m}$ with $\mu_1 \geq \cdots \geq \mu_m$ is *canonical*. Clearly every pattern has a canonical pattern with the same multiplicities. For example, the canonical pattern of 123223 is $1^3 2^2 3$.

We now define pattern probabilities. To be most general, we consider *mixed* distributions that assign probability to discrete elements and a continuous interval. For example, a distribution P may assign probability $p(a)$ to an element a, $p(b)$ to an element b, and $1 - p(a) - p(b)$ to the interval $[2, 3]$.

If P is sampled independently with replacement then

$$P(\overline{\psi}) \stackrel{\text{def}}{=} P(\{\overline{x} : \psi(\overline{x}) = \overline{\psi}\})$$

is the probability that the sample has pattern $\overline{\psi}$. For example, the distribution P above assigns to the pattern 121 probability

$$\begin{aligned} P(121) &= P(aba) + P(bab) + P(\{xyx : x \in \{a, b\}, y \in [2, 3]\}) \\ &= p^2(a)(1 - p(a)) + p^2(b)(1 - p(b)). \end{aligned}$$

Note that the pattern probability is determined by just the multiset of discrete probabilities, hence P can be identified with a vector in the monotone simplex

$$\mathcal{P} \stackrel{\text{def}}{=} \{(p_1, p_2, \dots) : p_1 \geq p_2 \geq \cdots \geq 0, \quad \sum p_i \leq 1\}.$$

We call $q \stackrel{\text{def}}{=} 1 - \sum p_i$, the *continuous part* of P. The *maximum-likelihood (PML) probability* of a pattern $\overline{\psi}$ is

$$\widehat{P}_{\overline{\psi}}(\overline{\psi}) \stackrel{\text{def}}{=} \max_{P \in \mathcal{P}} P(\overline{\psi}),$$

the highest probability assigned to $\overline{\psi}$ by any distribution, and its *maximum-likelihood (PML) distribution* $\widehat{P}_{\overline{\psi}}$ is the distribution achieving this highest probability. We let $\widehat{k} = \widehat{k}_{\overline{\psi}}$ denote the discrete support size of $\widehat{P}_{\overline{\psi}}$.

Observe that every distribution assigns the same probability to a pattern as it does to its canonical form. Hence the two have the same PML distribution. From now on we therefore consider without loss of generality only canonical patterns.

III. Results

A pattern is binary if, like 11122, it has $m = 2$. Theorem 11 in [12] shows that all binary patterns have $\widehat{k} = 2$, and the PML distribution can then be determined.

A pattern is *uniform* if, as in 121323, all multiplicities μ_i are equal. A pattern is *quasi-uniform* if the square of the difference between any two multiplicities is at most their sum, namely for all i, j, $(\mu_i - \mu_j)^2 \leq \mu_i + \mu_j$. For example, the pattern 111223 is quasi-uniform. Note that a binary pattern is quasi-uniform if $(\mu_1 - \mu_2)^2 \leq n$.

Theorem 11 in [12] shows also that all quasi-uniform binary patterns have PML $(\frac{1}{2}, \frac{1}{2})$. The following lemma extends this result to non-binary patterns when the underlying distribution is limited to support size m.

Lemma 1: If an m-symbol pattern is quasi-uniform then among all discrete distributions with support size m, its probability is maximized by the uniform distribution. ■

For example, the lemma implies that among all distributions over three elements, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ maximizes the probability of 111223.

The support-size restriction assumed in the lemma implies that it cannot be used to determine the PML distribution on its own. However, combined with other results that bound the support size it can be used to derive the PML distribution.

Theorem 6 in [12] states that

$$\widehat{k} \le m + \frac{m-1}{2^{\mu_m} - 2}. \tag{1}$$

In particular, all patterns with $\mu_m > \log(m+1)$ have PML distribution with support size m. Combined with the lemma, we obtain

Corollary 2: The PML distribution of a quasi-uniform pattern with $\mu_m > \log_2(m+1)$ is uniform over m symbols. ■

For example, the pattern 11111222333 is quasi-uniform and has $\mu_m = 3 > 2 = \log_2(m+1)$, hence the corollary yields the previously unknown PML distribution

$$\widehat{P}_{11111222333} = (1/3, 1/3, 1/3).$$

An important application of PML is to estimate the underlying distribution's support size $\widehat{k}$. Inequality (1) bounds the support size when the lowest multiplicity, μ_m, is at least 2. The next theorem upper bounds $\widehat{k}$ when $\mu_m = 1$ and all other multiplicities are at least 2, namely exactly one element appears once, for example as in the pattern 11122334. We call such patterns *unique-singleton*. We will later use this result to establish the PML distribution of ternary patterns.

Theorem 3: For unique-singleton patterns,

$$\widehat{k} \le 2(m-1). \quad ■$$

A pattern ψ is *1-uniform* if $\mu_i - \mu_j \le 1$ for all i, j, namely all multiplicities are within one from each other as in 1112233. As shown in [13], all 1-uniform patterns have a uniform PML distribution and can thus be evaluated.

As mentioned earlier, the simplest patterns are binary, and their PML distribution was derived in [12], showing in particular that all of them have $\widehat{k} = 2$. The next simplest patterns are ternary, and have $m = 3$. Three types of ternary patterns can be addressed by existing results.

1) Uniform ($1^r 2^r 3^r$). Of these, 123 has $\widehat{P} = ()$, and all others have $\widehat{P} = (1/3, 1/3, 1/3)$ [12].
2) 1-uniform ($1^r 2^r 3^{r-1}$ or $1^r 2^{r-1} 3^{r-1}$). Of these, 1123 has $\widehat{P} = (1/5, 1/5, 1/5, 1/5, 1/5)$, and all others have $\widehat{P} = (1/3, 1/3, 1/3)$ [13].
3) Skewed ($1^r 23$). Of these, 1123 is 1-uniform and addressed above, and all others have $\widehat{P} = (\frac{r}{r+2})$. This result is proved in [15].

It is easy to see that all ternary patterns not covered by these cases have at most one symbol appearing once, for example 111223 and 111122233. For all those, we show that the PML distribution has support size 3.

Theorem 4: All ternary patterns with at most one symbol appearing once have $\widehat{k} = 3$. ■

The theorem allows us to compute the PML distribution of all ternary patterns. Some follow by an easy combination of the theorem with Lemma 1.

Corollary 5: $\widehat{P}_{111223} = \widehat{P}_{1112223} = (1/3, 1/3, 1/3)$.

For more complex patterns, the PML distribution can be obtained by combining the theorem with the Kuhn-Tucker conditions.

Canonical $\overline{\psi}$	$\widehat{P}_{\overline{\psi}}$	Reference
1	any distribution	Trivial
11, 111, 111, ...	(1)	Trivial
12, 123, 1234, ...	()	Trivial
112, 1122, 1112, 11122, 111122	(1/2, 1/2)	[12]
11223, 112233, 1112233	(1/3, 1/3, 1/3)	[13]
111223, 1112223,	(1/3, 1/3, 1/3)	Corollary 5
1123, 1122334	(1/5, 1/5, ..., 1/5)	[12]
11234	(1/8, 1/8, ..., 1/8)	[13]
11123	(3/5)	[15]
11112	(0.7887.., 0.2113..)	[12]
111112	(0.8322.., 0.1678..)	[12]
111123	(2/3)	[15]
111234	(1/2)	[15]
112234	(1/6, 1/6, ..., 1/6)	[13]
112345	(1/13, ..., 1/13)	[13]
1111112	(0.857.., 0.143..)	[12]
1111122	(2/3, 1/3)	[12]
1112345	(3/7)	[15]
1111234	(4/7)	[15]
1111123	(5/7)	[15]
1111223	$\left(\frac{1}{\sqrt{7}}, \frac{\sqrt{7}-1}{2\sqrt{7}}, \frac{\sqrt{7}-1}{2\sqrt{7}}\right)$	Corollary 7
1123456	(1/19, ..., 1/19)	[13]
1112234	(1/5, 1/5, ..., 1/5)?	Conjectured

TABLE I
PML DISTRIBUTIONS OF ALL PATTERNS OF LENGTH ≤ 7

Corollary 6: For all ternary patterns with at most one symbol appearing once,

$$\widehat{P}_{1^{\mu_1}2^{\mu_2}3^{\mu_3}} = (p_1, p_2, p_3),$$

where p_1, p_2, p_3 are solutions to the following three polynomial equations,

$$p_1 + p_2 + p_3 = 1,$$
$$\sum \mu_{j_1} p_1^{\mu_{j_1}-1} p_2^{\mu_{j_2}} p_3^{\mu_{j_3}} = \sum \mu_{j_1} p_2^{\mu_{j_1}-1} p_3^{\mu_{j_2}} p_1^{\mu_{j_3}},$$
$$\sum \mu_{j_1} p_1^{\mu_{j_1}-1} p_2^{\mu_{j_2}} p_3^{\mu_{j_3}} = \sum \mu_{j_1} p_3^{\mu_{j_1}-1} p_1^{\mu_{j_2}} p_2^{\mu_{j_3}}.$$

where the summation is over all six permutations (j_1, j_2, j_3) of $(1, 2, 3)$. ■

For short patterns we can solve the equations in Corollary 6 and derive the PML distribution. An example is the following result.

Corollary 7: $\widehat{P}_{1111223} = \left(\frac{1}{\sqrt{7}}, \frac{1-\frac{1}{\sqrt{7}}}{2}, \frac{1-\frac{1}{\sqrt{7}}}{2}\right)$. ■

Combined with previously known results, the three PML distributions in Corollaries 5 and 7 yield the PML distributions of all but one pattern of length up to 7. The only exception is 1112234, which we conjecture to have PML $(1/5, 1/5, \ldots, 1/5)$ but have not been able to prove yet. The PML distributions of these patterns are shown in Table I along with references to where they were shown.

IV. PROOFS

For a probability distribution $P = (p_1, p_2, \ldots)$, let $\widetilde{P}_i = (p_1, \ldots, p_{i-1}, 0, p_{i+1}, \ldots)$ be the *sub-distribution* agreeing with P on all probabilities, except p_i, which is set to 0. Note that the probabilities in $\widetilde{P}_i$, including q, sum to $1 - p_i$, hence if $p_i > 0$ then $\widetilde{P}_i$ is not a distribution but a point inside the probability simplex $\mathcal{P}$. We let P_i, be normalized $\widetilde{P}_i$ so it is a

distribution. Note that the support size of P_i is one less than the support size of P. Similarly, let $\widetilde{P}_{i,j}$ be the sub-distribution obtained from P by setting p_i and p_j to 0, and let $P_{i,j}$ be its normalized version.

Similarly for a pattern $\overline{\psi} = 1^{\mu_1}2^{\mu_2}\cdots m^{\mu_m}$, let $\overline{\psi}_i \stackrel{\text{def}}{=} 1^{\mu_1}\cdots(i-1)^{\mu_{i-1}}i^{\mu_{i+1}}\cdots(m-1)^{\mu_m}$ be the pattern obtained by deleting all appearances of the ith symbol, and let $\overline{\psi}_{i,j}$ be the pattern obtained by deleting all appearances of ith and jth symbol.

The following hold for all $i \neq j \in \{1,\ldots,k\}$,

$$P(\overline{\psi}) = \widetilde{P}_i(\overline{\psi}) + \sum_{s=1}^{m} p_i^{\mu_s}\widetilde{P}_i(\overline{\psi}_s), \tag{2}$$

$$P(\overline{\psi}) = \widetilde{P}_{i,j}(\overline{\psi}) + \sum_{s=1}^{m}(p_i^{\mu_s}+p_j^{\mu_s})\widetilde{P}_{i,j}(\overline{\psi}_s) + \sum_{1\le s<t\le m}(p_i^{\mu_s}p_j^{\mu_t}+p_i^{\mu_t}p_j^{\mu_s})\widetilde{P}_{i,j}(\overline{\psi}_{s,t}). \tag{3}$$

Proof of Lemma 1: We show that if $P=(p_1,p_2,\ldots,p_m)$ is not uniform, namely $p_i > p_j$ for some $i<j$, then decreasing p_i and increasing p_j by the same small amount will increase $P(\overline{\psi})$. Using the fact that when support size of P is less than m, $P(\overline{\psi})=0$ in Equation (3),

$$P(\psi) = \sum_{1\le s<t\le m}(p_i^{\mu_s}p_j^{\mu_t}+p_i^{\mu_t}p_j^{\mu_s})\widetilde{P}_{i,j}(\overline{\psi}_{s,t}),$$

hence

$$\frac{\partial P(\overline{\psi})}{\partial p_i} = \sum_{1\le s<t\le m}(\mu_s p_i^{\mu_s-1}p_j^{\mu_t}+\mu_t p_i^{\mu_t-1}p_j^{\mu_s})\widetilde{P}_{i,j}(\overline{\psi}_{s,t}).$$

Since $\mu_s \ge \mu_t$ for $s<t$, it follows that

$$\frac{\partial P(\overline{\psi})}{\partial p_j} - \frac{\partial P(\overline{\psi})}{\partial p_i} = \sum_{1\le s<t\le m} f_{s,t}\cdot p_i^{\mu_t-1}\cdot p_j^{\mu_t-1}\cdot\widetilde{P}_{i,j}(\overline{\psi}_{s,t}),$$

where

$$\begin{aligned}
f_{s,t} =& \mu_s(p_ip_j)\left(p_j^{\mu_s-\mu_t-1}-p_i^{\mu_s-\mu_t-1}\right)\\
&+\mu_t\left(p_i^{\mu_s-\mu_t+1}-p_j^{\mu_s-\mu_t+1}\right)\\
=&(p_i-p_j)\cdot\\
&\left(\mu_t\left(p_i^{\mu_s-\mu_t}+p_j^{\mu_s-\mu_t}\right)-(\mu_s-\mu_t)\sum_{\alpha=1}^{\mu_s-\mu_t-1}p_i^{\alpha}p_j^{\mu_s-\mu_t-\alpha}\right)\\
\stackrel{(a)}{\ge}& (p_i-p_j)\left(p_i^{\mu_s-\mu_t}+p_j^{\mu_s-\mu_t}\right)\cdot\\
&\left(\mu_t-\frac{(\mu_s-\mu_t)(\mu_s-\mu_t-1)}{2}\right)\\
=&\frac{1}{2}\left(p_i^{\mu_s-\mu_t}+p_j^{\mu_s-\mu_t}\right)(p_i-p_j)\left((\mu_s+\mu_t)-(\mu_s-\mu_t)^2\right)\\
\stackrel{(b)}{\ge}& 0,
\end{aligned}$$

and (a) follows since for $1\le\alpha\le\mu_s-\mu_t-1$,

$$p_i^{\mu_s-\mu_t}+p_j^{\mu_s-\mu_t} \ge p_i^{\alpha}p_j^{\mu_s-\mu_t-\alpha}+p_j^{\alpha}p_i^{\mu_s-\mu_t-\alpha}.$$

Note that equality cannot hold simultaneously in both (a) and (b). Hence, $f_{s,t}>0$ for all $s<t$. Therefore

$$\frac{\partial P(\overline{\psi})}{\partial p_j} > \frac{\partial P(\overline{\psi})}{\partial p_i},$$

and decreasing p_i and increasing p_j by the same infinitesimal amount will increase $P(\overline{\psi})$. ■

For a distribution $P=(p_1,p_2,\ldots,p_k)$, denote the distribution obtained by combining the two smallest probabilities, p_{k-1} and p_k, by

$$P_{k-1\cup k} \stackrel{\text{def}}{=} (p_1,p_2,\ldots,p_{k-2},p_{k-1}+p_k).$$

Proof of Theorem 3: Theorem 2 of [12] states that PML distribution is discrete whenever at most one symbol appears once. Hence we consider only discrete distributions. We show that for any distribution P with support size $k>2m-2$, $P_{k-1\cup k}(\overline{\psi}) > P(\overline{\psi})$. This implies that $\widehat{k}\le 2(m-1)$.

Let $k\ge 2m-1$. We will prove that for any distribution P with support size k, $P(\overline{\psi}) < P_{k-1\cup k}(\overline{\psi})$. By taking $i=k$ and $j=k-1$ in Equation (3) we obtain

$$P(\overline{\psi}) = \widetilde{P}_{k-1,k}(\overline{\psi}) + \sum_{i=1}^{m}(p_k^{\mu_i}+p_{k-1}^{\mu_i})\widetilde{P}_{k-1,k}(\overline{\psi}_i) + \sum_{1\le i<j\le m}(p_k^{\mu_j}p_{k-1}^{\mu_i}+p_k^{\mu_i}p_{k-1}^{\mu_j})\widetilde{P}_{k-1,k}(\overline{\psi}_{i,j}).$$

Taking $i=k-1$ in Equation (2) in the distribution $P_{k-1\cup k}$,

$$\begin{aligned}
P_{k-1\cup k}(\overline{\psi}) &= \widetilde{P}_{k-1,k}(\overline{\psi}) + \sum_{i=1}^{m}(p_k+p_{k-1})^{\mu_i}\widetilde{P}_{k-1,k}(\overline{\psi}_i)\\
&\stackrel{(a)}{\ge} \widetilde{P}_{k-1,k}(\overline{\psi}) + \sum_{i=1}^{m}(p_k^{\mu_i}+p_{k-1}^{\mu_i})\widetilde{P}_{k-1,k}(\overline{\psi}_i)\\
&\quad+\sum_{i=1}^{m-1}(p_kp_{k-1}^{\mu_i-1}+p_k^{\mu_i-1}p_{k-1})\widetilde{P}_{k-1,k}(\overline{\psi}_i),
\end{aligned}$$

where (a) follows since for $\mu\ge 2$,

$$(p_k+p_{k-1})^{\mu} \ge p_{k-1}^{\mu}+p_{k-1}^{\mu-1}p_k+p_{k-1}p_k^{\mu-1}+p_{k-1}^{\mu}.$$

To prove $P(\overline{\psi}) < P_{k-1\cup k}(\overline{\psi})$ it suffices to show that

$$\sum_{i=1}^{m-1}(p_kp_{k-1}^{\mu_i-1}+p_k^{\mu_i-1}p_{k-1})\widetilde{P}_{k-1,k}(\overline{\psi}_i) \ge \textstyle\sum_{1\le i<j\le m}(p_k^{\mu_j}p_{k-1}^{\mu_i}+p_k^{\mu_i}p_{k-1}^{\mu_j})\widetilde{P}_{k-1,k}(\overline{\psi}_{i,j}).$$

Claim 1:

$$\widetilde{P}_{k-1,k}(\overline{\psi}_i) \ge \frac{(k-m)\sum_{j\neq i}^{m}p_{k-1}^{\mu_j}\widetilde{P}_{k-1,k}(\overline{\psi}_{i,j})}{(m-1)}.$$

Proof:

$$\begin{aligned}
\widetilde{P}_{k-1,k}(\overline{\psi}_i) &= \sum_{s=1}^{k-2}p_s^{\mu_j}\widetilde{P}_{k-1,k,s}(\overline{\psi}_{i,j})\\
&\stackrel{(a)}{\ge} \sum_{s=1}^{k-2}p_{k-1}^{\mu_j}\widetilde{P}_{k-1,k,s}(\overline{\psi}_{i,j})\\
&\stackrel{(b)}{=} p_{k-1}^{\mu_j}((k-2)-(m-2))\widetilde{P}_{k-1,k}(\overline{\psi}_{i,j}),
\end{aligned}$$

where (a) follows since p_i's are non decreasing and (b) follows since each term in the polynomial expansion of $\widehat{P}_{k-1,k}(\overline{\psi}_{i,j})$ appears in all but $m-2$ summands. Summing over all j's not equal to i yields the claim.

Using Claim 1, whenever $k-m \geq m-1$, we will show that for any i and j, the coefficient of $\widehat{P}_{k-1,k}(\overline{\psi}_{i,j})$ is larger for $P_{k-1\cup k}$. This is equivalent to showing that

$$(p_k p_{k-1}^{\mu_i-1} + p_k^{\mu_i-1} p_{k-1}) p_{k-1}^{\mu_j} \geq (p_k^{\mu_j} p_{k-1}^{\mu_i} + p_k^{\mu_i} p_{k-1}^{\mu_j}),$$

which follows since $p_{k-1} \geq p_k$. ■

Proof of Theorem 4: If $\mu_1 = 2$, then $\overline{\psi} = 11223$ or 112233 is 1-uniform, and [13] implies $\widehat{k} = 3$. Assume then that $\mu_1 \geq 3$. If $\mu_3 \geq 3$, then Equation (1) shows that $\widehat{k} = 3$. Thus we assume $\mu_3 \leq 2$. Theorem 3 implies $\widehat{k} \leq 4$. Let $P = (p_1, p_2, p_3, p_4)$ be any discrete distribution with support size 4. We show that $P_{3\cup 4} = (p_1, p_2, p_3 + p_4)$ assigns larger probability to any pattern with $\mu_1 \geq 3$, $\mu_2 \geq 2$ and $\mu_3 \leq 2$. We provide a sketch of the proof for $\mu_3 = 1$. An identical argument holds for $\mu_3 = 2$.

Proceeding as in Theorem 3 and using the following facts which hold for $\mu_2 \geq 2$ and $\mu_1 \geq 3$,

$$\begin{aligned}(p_k + p_{k-1})^{\mu_2} &\geq p_k^{\mu_2} + p_{k-1}^{\mu_2-1} p_k + p_{k-1} p_k^{\mu_2-1} + p_{k-1}^{\mu_2} \\ (p_k + p_{k-1})^{\mu_1} &\geq p_k^{\mu_1} + 3(p_{k-1}^{\mu_1-1} p_k + p_{k-1} p_k^{\mu_1-1}) + p_{k-1}^{\mu_1},\end{aligned}$$

it suffices to show that

$$\begin{aligned}&(p_1^{\mu_1} p_2 + p_2^{\mu_1} p_1)(p_3^{\mu_2-1} p_4 + p_4^{\mu_2-1} p_3) \\ &\quad + 3(p_3^{\mu_1-1} p_4 + p_4^{\mu_1-1} p_3)(p_1^{\mu_2} p_2 + p_2^{\mu_2} p_1) \\ &\geq (p_1 + p_2)(p_3^{\mu_1} p_4^{\mu_2} + p_4^{\mu_1} p_3^{\mu_2}) \\ &\quad + (p_1^{\mu_1} + p_2^{\mu_1})(p_3^{\mu_2} p_4 + p_4^{\mu_2} p_3) + (p_2^{\mu_2} + p_1^{\mu_2})(p_3^{\mu_1} p_4 + p_4^{\mu_1} p_3).\end{aligned}$$

This can be verified by expanding and suitably pairing the terms and using the fact that $\mu_3 \geq \mu_4$. ■

This can be used to find the PML distribution for all the patterns with three distinct symbols. Skewed [15] and 1-uniform patterns [13] have been proved. The remaining patterns can be solved by the method mentioned here.

We now find the exact PML distribution for some short patterns, using the tools developed.

Proof of Corollary 5: By Lemma 4, the support size of the PML distribution is 3. The multiplicities satisfy the condition of Lemma 1, thus proving the theorem. ■

Proof of Corollary 6: By Theorem 4 we know that the PML distribution is of the form (p_1, p_2, p_3). Kuhn-Tucker conditions state that p_i's satisfy

$$\frac{\partial P(\overline{\psi})}{\partial p_1} = \frac{\partial P(\overline{\psi})}{\partial p_2} = \frac{\partial P(\overline{\psi})}{\partial p_3}.$$

Using Equation (2), we get the conditions mentioned. ■

Proof of Corollary 7: By Theorem 4, the support size of the PML distribution is 3. Let $\widehat{P}_{1111223} = (p_1, p_2, p_3)$. We show that two of the p_i's have same value.

Let $S_1 = p_1 + p_2 + p_3$, $S_2 = p_1^2 + p_2^2 + p_3^2$, $T_2 = p_1p_2 + p_3p_1 + p_2p_3$, and $T_3 = p_1p_2p_3$. By Corollary 6,

$$\frac{\partial P(\overline{\psi})}{\partial p_1} - \frac{\partial P(\overline{\psi})}{\partial p_2} = 0,$$

which for $p_1 \neq p_2$ yields

$$4T_3T_2 - 2T_3(p_1^2 + p_2^2 + p_1p_2) - (p_1 + p_2)p_3^2 S_2 = 0.$$

Similarly, if $p_1 \neq p_3$,

$$4T_3T_2 - 2T_3(p_1^2 + p_3^2 + p_1p_3) - (p_1 + p_3)p_2^2 S_2 = 0$$

Subtracting, we obtain

$$(p_2 - p_3)(2S_1T_3 - S_2T_2) = 0.$$

Using the arithmetic-geometric mean inequality it is easy to see that $S_2T_2 > 2S_1T_3$.

This would mean that $p_2 = p_3$, hence two of p_i's are the same. Let the PML distribution be $\widehat{P}_{1111223} = \left(p, \frac{1-p}{2}, \frac{1-p}{2}\right)$ and we have to maximize

$$\widehat{P}(1111223) = \frac{p(1-p)^3(1-p-p^2+9p^3)}{32}.$$

Differentiating and equating to 0, $(1-7p^2)(3p+1)^2 = 0$. Thus, the only maximizing value of p is $\frac{1}{\sqrt{7}}$. ■

REFERENCES

[1] R. Fisher, A. Corbet, and C. Williams, "The relation between the number of species and the number of individuals in a random sample of an animal population," *Journal of Animal Ecology*, vol. 12, pp. 42–48, 1943.

[2] I. Good and G. Toulmin, "The number of new species and the increase in population coverage when the sample is increased," *Biometrika*, vol. 43, no. 1, pp. 45–63, 1956.

[3] B. Efron and R. Thisted, "Estimating the number of unseen species: How many words did Shakespeare know," *Biometrika*, vol. 63, pp. 435–447, 1976.

[4] J. Bunge and M. Fitzpatrick, "Estimating the number of species: a review," *Journal of the American Statistical Association*, vol. 88, pp. 364–373, 1993.

[5] P. Laplace, *Philosphical essays on probabilities*, Translated by A. Dale from the 5th (1825) ed. Springer Verlag, New York, 1995.

[6] I. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3/4, pp. 237–264, December 1953.

[7] D. McAllester and R. Schapire, "On the convergence rate of Good Turing estimators," in *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

[8] G. Shamir, "Universal lossless compression with unknown alphabets—the average case," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 4915–4944, November 2006.

[9] G. M. Gemelos and T. Weissman, "On the entropy rate of pattern processes," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3994–4007, 2006.

[10] B. M. Jedynak and S. Khudanpur, "Maximum likelihood set for estimating a probability mass function," *Neural Computation*, vol. 17, pp. 1–23, 2005.

[11] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "A better good-turing estimator for sequence probabilities," *CoRR*, vol. abs/0704.1455, 2007.

[12] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004.

[13] ——, "Pattern maximum likelihood: existence and properties," In preparation, 2009.

[14] A. Orlitsky, Sajama, N. Santhanam, K. Viswanathan, and J. Zhang, "Pattern maximum likelihood: computation and experiments," In preparation, 2009.

[15] A. Orlitsky and S. Pan, "The maximum likelihood probability of skewed patterns," Accepted at IEEE International Symposium on Information Theory, 2009.