# A Phase Transition for the Uniform Distribution in the Pattern Maximum Likelihood Problem

Winston Fernandes and Navin Kashyap
Dept. of Electrical Communication Engineering
Indian Institute of Science, Bangalore, India
Email: {winston,nkashyap}@ece.iisc.ernet.in

***Abstract*—In this paper, we consider the setting of the pattern maximum likelihood (PML) problem studied by Orlitsky et al. We present a well-motivated heuristic algorithm for deciding the question of when the PML distribution of a given pattern is uniform. The algorithm is based on the concept of a "uniform threshold". This is a threshold at which the uniform distribution exhibits an interesting phase transition in the PML problem, going from being a local maximum to being a local minimum.**

## I. Introduction

Consider the problem where we are given a sequence of $n$ i.i.d. samples from a fixed but unknown underlying probability distribution $\mathbf{p}$, and we are required to estimate the multiset of probabilities in $\mathbf{p}$. In particular, we are not required to determine the correspondence between the symbols of the underlying alphabet and the probabilities in the multiset. Such a problem arises naturally in the context of universal compression of large-alphabet sources [1], and has several other applications, for example, population estimation from a small number of samples [2].

The *pattern* $\boldsymbol{\psi}$ or $\boldsymbol{\psi}(\mathbf{x}^n)$ of a sequence $\mathbf{x}^n = x_1, \ldots, x_n$ is a data structure that keeps track of the order of occurrence and the multiplicities of the distinct symbols in the sequence $\mathbf{x}^n$; for a precise definition, see Section II. The *pattern maximum likelihood (PML) distribution* of a pattern $\boldsymbol{\psi}$ is the multiset of probabilities that maximizes the probability of observing a sequence with pattern $\boldsymbol{\psi}$. It has been observed that for a sequence $\mathbf{x}^n$ sampled from an unknown underlying probability distribution $\mathbf{p}$, the PML distribution of $\boldsymbol{\psi}(\mathbf{x}^n)$ is a good estimate of the multiset of probabilities in $\mathbf{p}$ in situations where $n$ is much smaller than the support size of $\mathbf{p}$ [3], [4].

However, the problem of determining the PML distribution (henceforth termed the "PML problem") of a given pattern $\boldsymbol{\psi}$ appears to be computationally hard. This problem has been analytically solved for all patterns of length up to 14 [5]–[9]. Algorithms for approximating the PML distribution exist in the literature [10], [11].

In this paper, we are motivated by the possibly simpler question of determining when the PML distribution of a pattern is a uniform distribution. While we are unable to give a complete answer to this question, we are able to prove a remarkable phase transition phenomenon. For $k \in \mathbb{Z}_+$, let $U_k$ denote the uniform distribution on $k$ symbols. Given a pattern $\boldsymbol{\psi}$, we can explicitly compute a quantity $\Theta(\boldsymbol{\psi})$ such that for all $k < \Theta(\boldsymbol{\psi})$, $U_k$ is a *local maximum* for the PML problem, among all distributions $\mathbf{p}$ with support size $k$; and for all $k > \Theta(\boldsymbol{\psi})$, $U_k$ is a *local minimum*. We call $\Theta(\boldsymbol{\psi})$ the *uniform threshold* of $\boldsymbol{\psi}$.

Based on the uniform threshold, we propose a heuristic for classifying a pattern as having a uniform PML distribution (a *uniform pattern*) or a non-uniform PML distribution (a *non-uniform pattern*). For a pattern $\boldsymbol{\psi}$ that is classified as having a uniform PML distribution, the algorithm also gives the support size $k$ of the distribution $U_k$ that maximizes, among all uniform distributions, the probability of observing a sequence with pattern $\boldsymbol{\psi}$. If the classification by the algorithm is correct, then this $U_k$ is the true PML distribution for $\boldsymbol{\psi}$; otherwise, it works as a good approximation of the PML distribution. For a non-uniform pattern, the algorithm from [11] can be used to find an approximate PML distribution. We present analytical and experimental evidence in support of the validity of our proposed pattern classification algorithm.

The rest of this paper is organized as follows. We provide the necessary definitions and notation in Section II. In Section III, we state and prove the phase transition phenomenon mentioned above, and in doing so, give an explicit expression for the uniform threshold. In Section IV, we study a quantity called the "optimal uniform support size". The relationship between the uniform threshold and the optimal uniform support size forms the basis of our proposed pattern classification algorithm, described in Section V. Section VI contains some experimental results in support of our algorithm. The paper ends in Section VII with a conjecture.

## II. Definitions and Notation

Given a sequence $\mathbf{x}^n = x_1, \ldots, x_n$ over some alphabet, the *pattern* of $\mathbf{x}^n$ is the sequence $\boldsymbol{\psi} = \psi_1, \psi_2, \ldots, \psi_n$ obtained by replacing each $x_j$ by the order of its first occurrence in $\mathbf{x}^n$ [11]. More precisely, for each symbol $x$ occurring in $\mathbf{x}^n$, let $\nu(x)$ denote the number of distinct symbols seen in the *shortest* prefix of $\mathbf{x}^n$ that ends in the symbol $x$. Then, $\psi_j = \nu(x_j)$ for $j = 1, 2, \ldots, n$. The pattern $\boldsymbol{\psi}(\mathbf{x}^n)$ is defined to have *length* $n$ and *size* $m$, where $m$ is the number of distinct symbols in $\mathbf{x}^n$.

For example, the word "sleepless" has pattern 123342311, which is of length 9 and size 4. We will canonically represent a pattern $\boldsymbol{\psi}$ as $1^{\mu_1}2^{\mu_2}\dots m^{\mu_m}$, where $\mu_j$ is the *multiplicity* of the symbol $j$, i.e., the number of times $j$ appears, in $\boldsymbol{\psi}$. Note that $\mu_1+\cdots+\mu_m = n$. The pattern $\boldsymbol{\psi}$ in our example has canonical form $1^3 2^2 3^3 4$.

We now define pattern probabilities. Let $\mathbf{p} = (p(x))_{x\in\mathcal{X}}$ be a probability distribution over a discrete set $\mathcal{X}$. A sequence $\mathbf{x}^n$ obtained by taking $n$ i.i.d. samples from $\mathbf{p}$ will have probability given by $P(\mathbf{x}^n) \triangleq \prod_{j=1}^n p(x_j)$. The probability of observing a sequence with pattern $\boldsymbol{\psi}$ of length $n$ is then given by

$$P(\boldsymbol{\psi};\mathbf{p}) \triangleq \sum_{\mathbf{x}^n:\boldsymbol{\psi}(\mathbf{x}^n)=\boldsymbol{\psi}} P(\mathbf{x}^n).$$

Clearly, all patterns $\boldsymbol{\psi}$ with the same canonical form $1^{\mu_1}2^{\mu_2}\dots m^{\mu_m}$ will have the same pattern probability $P(\boldsymbol{\psi};\mathbf{p})$. If $\mathbf{p}$ is the uniform distribution $U_k$, then it is readily verified that for any pattern $\boldsymbol{\psi}$ of length $n$ and size $m \le k$, we have $P(\boldsymbol{\psi};U_k) = P_{\mathsf{u}}(n,m;k)$, where

$$P_{\mathsf{u}}(n,m;k) \triangleq k(k-1)(k-2)\cdots(k-m+1)(1/k)^n. \quad (1)$$

The *PML probability* of a pattern $\boldsymbol{\psi}$ is defined as

$$P^{\mathrm{PML}}(\boldsymbol{\psi}) \triangleq \max_{\mathbf{p}} P(\boldsymbol{\psi};\mathbf{p})$$

the maximum being taken over all discrete distributions $\mathbf{p}$. Any distribution that attains the maximum above is called a *PML distribution* of $\boldsymbol{\psi}$, denoted by $\mathbf{p}^{\mathrm{PML}}(\boldsymbol{\psi})$. For the purposes of this paper, we will assume that the maximum is indeed attained by some discrete distribution $\mathbf{p}$.[1]

## III. Phase Transition at the Uniform Threshold

In this section, we detail the phase transition phenomenon briefly described in the introduction. For $k \in \mathbb{Z}_+$, let $[k]$ denote the set $\{1,2,\dots,k\}$, and let $\Pi_{[k]}$ denote the simplex of probability distributions on $[k]$:

$$\Pi_{[k]} = \{\mathbf{p} = (p_1,\dots,p_k) : p_j \ge 0\ \forall j,\ \sum_{j=1}^k p_j = 1\}.$$

Let $\boldsymbol{\psi}$ be a pattern having canonical form $1^{\mu_1}2^{\mu_2}\dots m^{\mu_m}$. For any $\mathbf{p} \in \Pi_{[k]}$ with $k \ge m$, we can write

$$P(\boldsymbol{\psi};\mathbf{p}) = \sum_{\sigma}\prod_{i=1}^m p_{\sigma(i)}^{\mu_i}, \quad (2)$$

where the summation runs over all one-to-one maps $\sigma : [m] \to [k]$. Let $\pi_{[k]}^{\boldsymbol{\psi}} : \Pi_{[k]} \to [0,1]$ be the function defined by the mapping $\mathbf{p} \mapsto P(\boldsymbol{\psi};\mathbf{p})$. We then have the following phase transition phenomenon.

**Theorem 1.** *For a pattern $\boldsymbol{\psi}$ having canonical form $1^{\mu_1}2^{\mu_2}\dots m^{\mu_m}$, with $n = \mu_1+\cdots+\mu_m$, define*

$$\Theta(\boldsymbol{\psi}) = \frac{n^2-n}{\sum_{i=1}^m \mu_i^2 - n}. \quad (3)$$

*Then, for $m \le k < \Theta(\boldsymbol{\psi})$, the uniform distribution $U_k$ is a local maximum of the function $\pi_{[k]}^{\boldsymbol{\psi}}$, and for $k > \Theta(\boldsymbol{\psi})$, $U_k$ is a local minimum.*

*Proof:* The proof approach is based on that of Theorem 20 in [12]. Let $\mathbf{p} = U_k$, so that $\mathbf{p}$ is in the interior of the simplex $\Pi_{[k]}$. Pick an arbitrary direction $\boldsymbol{\xi} \in \mathbb{R}^k \setminus \{\mathbf{0}\}$ such that for all $t$ within a sufficiently small interval around 0, the point $\mathbf{p}(t) = \mathbf{p} + t\,\boldsymbol{\xi}$ continues to lie within $\Pi_{[k]}$. Note that this implies that $\sum_{j=1}^k \xi_j = 0$. Consider the function $g(t) = P(\boldsymbol{\psi};\mathbf{p}(t))$. We will show that $g'(0) = 0$, $g''(0) < 0$ if $k < \Theta(\boldsymbol{\psi})$, and $g''(0) > 0$ if $k > \Theta(\boldsymbol{\psi})$. Since the direction $\boldsymbol{\xi}$ is arbitrary, this suffices to prove the theorem.

Now, from (2), $g(t)$ is expressible as $\sum_\sigma g_\sigma(t)$, where $g_\sigma(t) = \prod_{i=1}^m (p_{\sigma(i)} + t\,\xi_{\sigma(i)})^{\mu_i}$. Differentiation, together with the fact that $p_j = \frac{1}{k}$ for all $j$, yields $g'_\sigma(0) = \frac{1}{k^{n-1}}\sum_{i=1}^m \mu_i \xi_{\sigma(i)}$. Hence,

$$g'(0) = \sum_\sigma g'_\sigma(0) = \frac{1}{k^{n-1}}\sum_{i=1}^m \mu_i \sum_\sigma \xi_{\sigma(i)}.$$

For any fixed $i \in [m]$, the inner summation $\sum_\sigma \xi_{\sigma(i)}$ can be evaluated as follows. As $\sigma$ ranges over all one-to-one maps from $[m]$ to $[k]$, for each $j \in [k]$, $\sigma(i)$ takes the value $j$ exactly $\frac{(k-1)!}{(k-m)!}$ times. Hence, $\sum_\sigma \xi_{\sigma(i)} = \frac{(k-1)!}{(k-m)!}\sum_{j=1}^k \xi_j = 0$ by choice of $\boldsymbol{\xi}$. Thus, $g'(0) = 0$.

Next, we compute $g''(0) = \sum_\sigma g''_\sigma(0)$. Straightforward computations yield

$$g''_\sigma(0) = \frac{1}{k^{n-2}}\left[\left(\sum_{i=1}^m \mu_i \xi_{\sigma(i)}\right)^2 - \sum_{i=1}^m \mu_i \xi_{\sigma(i)}^2\right].$$

Re-write the term within square brackets as

$$\sum_{i=1}^m \mu_i(\mu_i-1)\xi_{\sigma(i)}^2 + \sum_{(i,\ell):i\ne\ell} \mu_i\mu_\ell \xi_{\sigma(i)}\xi_{\sigma(\ell)}.$$

Summing over all one-to-one maps $\sigma : [m] \to [k]$, we obtain

$$\sum_{i=1}^m \mu_i(\mu_i-1)\sum_\sigma \xi_{\sigma(i)}^2 + \sum_{(i,\ell):i\ne\ell} \mu_i\mu_\ell \sum_\sigma \xi_{\sigma(i)}\xi_{\sigma(\ell)}.$$

As above, $\sum_\sigma \xi_{\sigma(i)}^2 = \frac{(k-1)!}{(k-m)!}\sum_{j=1}^k \xi_j^2$. Similarly, for $i \ne \ell$, $\sum_\sigma \xi_{\sigma(i)}\xi_{\sigma(\ell)} = \frac{(k-2)!}{(k-m)!}\sum_{(s,t)\in[k]^2:s\ne t} \xi_s\xi_t$. We also have $0 = \left(\sum_{j=1}^k \xi_j\right)^2$, from which we obtain $\sum_{j=1}^k \xi_j^2 = -\sum_{(s,t)\in[k]^2:s\ne t}\xi_s\xi_t$. Hence, $\sum_\sigma \xi_{\sigma(i)}\xi_{\sigma(\ell)} = -\frac{(k-2)!}{(k-m)!}\sum_{j=1}^k \xi_j^2$. Putting it all together, we find that

$$g''(0) = C(\boldsymbol{\xi})\left[(k-1)\sum_{i=1}^m \mu_i(\mu_i-1) - \sum_{(i,\ell)\in[m]^2:i\ne\ell}\mu_i\mu_\ell\right],$$

where $C(\boldsymbol{\xi}) = \frac{1}{k^{n-2}}\left(\sum_{j=1}^k \xi_j^2\right)\frac{(k-2)!}{(k-m)!}$ is a positive factor. Further simplification using the fact that $\sum_{i=1}^m \mu_i = n$

[1] In general, to guarantee that the maximum is always attained, we must allow "mixed" distributions; see [1], [2].

yields

$$g''(0) = C(\boldsymbol{\xi}) \left[ k \left( \sum_{i=1}^{m} \mu_i^2 - n \right) - (n^2 - n) \right],$$

from which the desired result follows. ∎

The quantity $\Theta(\boldsymbol{\psi})$ will be called the *uniform threshold* for the pattern $\boldsymbol{\psi}$. Note that the uniform threshold is finite iff $\boldsymbol{\psi} \neq 123\ldots n$.

The following is an immediate consequence of Theorem 1.

**Corollary 1.1.** *A necessary condition for the PML distribution of a pattern $\boldsymbol{\psi}$ to be uniform is that $\Theta(\boldsymbol{\psi}) \geq m$.*

For example, if $\boldsymbol{\psi}$ is such that $\mu_1 > 1$ and $\mu_i = 1$ for $2 \leq i \leq m$, then $\Theta(\boldsymbol{\psi}) = \frac{n^2-n}{\mu_1^2-\mu_1}$ and $m = n - (\mu_1 - 1)$. In this case, it may be verified that $\Theta(\boldsymbol{\psi}) \geq m$ iff $\mu_1 \leq 1+\sqrt{n}$. In the next section, we give a stronger necessary condition for a pattern to have a uniform PML distribution, based on which we develop a pattern classification algorithm.

We end this section with an observation about the convergence behaviour of the uniform threshold. Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. samples drawn from a distribution $\mathbf{p}$ supported on a finite set $[k]$. For $j \in [k]$, let $f_j(\mathbf{X}^n)$ denote the empirical frequency of the symbol $j$ in $\mathbf{X}^n \triangleq X_1, X_2, \ldots, X_n$. By the strong law of large numbers, $f_j(\mathbf{X}^n)$ converges to $p_j$ for all $j \in [k]$, almost surely. Now, consider the uniform threshold for the pattern $\boldsymbol{\psi}(\mathbf{X}^n)$, which can be expressed as

$$\Theta\big(\boldsymbol{\psi}(\mathbf{X}^n)\big) = \frac{1 - \frac{1}{n}}{\sum_{i=1}^{m} (\frac{\mu_i}{n})^2 - \frac{1}{n}}.$$

The sum $\sum_{i=1}^{m} (\frac{\mu_i}{n})^2$ is the same as $\sum_{j=1}^{k} f_j^2$, which converges almost surely to $\sum_{j=1}^{n} p_j^2$. Hence, the (random) sequence of uniform thresholds $\Theta\big(\boldsymbol{\psi}(\mathbf{X}^n)\big)$, $n = 1, 2, \ldots$, converges almost surely to $\frac{1}{\sum_{j=1}^{k} p_j^2}$.

## IV. Optimal Uniform Support Size

A stronger necessary condition than that in Corollary 1.1 can be obtained from the following observation. For a pattern $\boldsymbol{\psi}$ to have a uniform distribution $U_{k^*}$ as its PML distribution, it must be the case that

(a) $U_{k^*}$ maximizes the pattern probability $P(\boldsymbol{\psi}; U_k)$ among all uniform distributions $U_k$; and
(b) $k^* \leq \Theta(\boldsymbol{\psi})$.

With this in mind, we define the *optimal uniform support size* of a pattern $\boldsymbol{\psi}$ of length $n$ and size $m$ to be an integer $k$ that achieves the maximum in

$$\max_{k:k \geq m} P_{\mathsf{u}}(n, m; k), \tag{4}$$

where $P_{\mathsf{u}}(n, m; k)$ is as defined in (1). The optimal uniform support size will be denoted by $k_{\mathsf{u}}^*(\boldsymbol{\psi})$ or $k_{\mathsf{u}}^*(n, m)$. We then have the following corollary to Theorem 1.

**Corollary 1.2.** *A necessary condition for the PML distribution of a pattern $\boldsymbol{\psi}$ to be uniform is that $\Theta(\boldsymbol{\psi}) \geq k_{\mathsf{u}}^*(\boldsymbol{\psi})$.*

A few clarifying remarks about the definition of $k_{\mathsf{u}}^*(\boldsymbol{\psi})$ are in order. If $\boldsymbol{\psi}$ is such that $m = n = 1$ (i.e., $\boldsymbol{\psi} = 1$), then $P(\boldsymbol{\psi}; \mathbf{p}) = 1$ for all distributions $\mathbf{p}$. Thus, $k_{\mathsf{u}}^*(\boldsymbol{\psi})$ can be any $k \in \mathbb{Z}_+$ in this case. If $m = n > 1$, then $P_{\mathsf{u}}(m, m; k) = \prod_{i=1}^{m-1}(1 - \frac{i}{k})$, which strictly increases with $k$, and the maximum in (4) is not attained for any finite $k$. In this case, we simply define $k_{\mathsf{u}}^*(m, m) = \infty$. The following proposition covers all other cases.

**Proposition 2.** *Let $\boldsymbol{\psi}$ be a pattern of length $n$ and size $m < n$. Then, the maximum in (4) is attained by a unique integer $k \geq m$, i.e., $k_{\mathsf{u}}^*(n, m)$ is unique. Furthermore, $P_{\mathsf{u}}(n, m; k)$ is a strictly increasing function of $k$ for $m \leq k \leq k_{\mathsf{u}}^*(n, m)$, and is a strictly decreasing function of $k$ for $k \geq k_{\mathsf{u}}^*(n, m)$.*

*Proof:* Consider the function

$$f(x) = x(x-1)\cdots(x-m+1)(1/x)^n,$$

and note, from (1), that $P_{\mathsf{u}}(n, m; k) = f(k)$ for $k \geq m$. Taking the derivative, we obtain

$$f'(x) = \frac{f(x)}{x} \left[ \frac{x}{x-1} + \frac{x}{x-2} + \cdots + \frac{x}{x-m+1} - (n-1) \right]$$

Since $f(x)/x$ is positive for $x \geq m$, the sign of $f'(x)$ is determined by the factor within square brackets above, which we will denote by $\varphi(x)$. Note that we can write

$$\varphi(x) = \frac{1}{x-1} + \frac{2}{x-2} + \cdots + \frac{m-1}{x-(m-1)} + (m-n). \tag{5}$$

Thus, $\varphi(x)$ is strictly decreasing in $x$, and since $m < n$, $\varphi(x) < 0$ for all sufficiently large $x$.

Therefore, we have one of two cases:

(a) $\varphi(x) < 0$, and hence, $f'(x) < 0$ for all $x \geq m$;
(b) there exists a unique $x^* \geq m$ such that $\varphi(x) = f'(x) = 0$; for $m \leq x < x^*$, we have $\varphi(x) > 0$, and hence, $f'(x) > 0$; and for $x > x^*$, we have $\varphi(x) < 0$, and hence, $f'(x) < 0$.

In the first case, $f(k)$ is strictly decreasing in the range $k \geq m$, with a unique maximum attained at $k = m$; thus, $k_{\mathsf{u}}^*(n, m) = m$ here.

In the second case, $f(x)$ strictly increases up to $x^*$, and thereafter, strictly decreases. In particular, $x^*$ is the unique maximum of $f(x)$ in the range $x \geq m$. Therefore, if $x^*$ is an integer, then $k_{\mathsf{u}}^*(n, m) = x^*$. Otherwise, $f(k) = P_{\mathsf{u}}(n, m; k)$ potentially has two maxima (among integers $k \geq m$): $k = \lfloor x^* \rfloor$ and $k + 1 = \lceil x^* \rceil$. However, note that if we set $f(k) = f(k+1)$, and solve for $m$, we obtain

$$m = (k+1) - \frac{k^n}{(k+1)^{n-1}}.$$

For positive integers $k, m, n$, this is possible iff $m = n = 1$, which is not allowed by the assumption that $n > m$.

Therefore, once again, $f(k) = P_{\mathsf{u}}(n, m; k)$ has a unique maximum among the integers $k \geq m$. ∎

The following corollary will be useful later.

**Corollary 2.1.** *For any fixed $m$, we have $k^*_u(n, m) = m$ for all $n > m + \sum_{j=1}^{m-1} \frac{j}{m-j}$.*

*Proof:* From (5), we see that for $n > m + \sum_{j=1}^{m-1} \frac{j}{m-j}$, we have $\varphi(m) < 0$, and hence, $\varphi(x) < 0$ for all $x \geq m$. Thus, case (a) in the proof of Proposition 2 applies, and hence, $k^*_{\mathsf{u}}(n, m) = m$. ∎

From Proposition 2, it is clear that when $m < n$, the optimal uniform support size can be computed by stepping through the integers $k \geq m$ until the pattern probability $P_{\mathsf{u}}(n, m; k)$ does not increase. This is formalized in Algorithm 1. The condition in the while loop of the algorithm is equivalent to $P_{\mathsf{u}}(n, m; k) < P_{\mathsf{u}}(n, m; k + 1)$.

**Algorithm 1**

**procedure** OPTIMAL UNIFORM SUPPORT SIZE($n, m$)
  $k := m$
  **while** $\left(\frac{k-m+1}{k^n} < \frac{1}{(k+1)^{n-1}}\right)$ **do**
    $k \leftarrow k + 1$
  **end while**
  **return** $k$
**end procedure**

The limiting behaviour of the optimal uniform support size differs from that of the uniform threshold discussed at the end of Section III. Again, let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. samples drawn from a distribution $\mathbf{p}$ with support $[k]$. From the strong law of large numbers, it follows that the size $m$ of the patterns $\boldsymbol{\psi}(\mathbf{X}^n)$ converges to $k$ almost surely. In other words, almost surely, the size of the patterns $\boldsymbol{\psi}(\mathbf{X}^n)$ equals $k$ for all sufficiently large $n$. Therefore, from Corollary 2.1, we obtain that $k^*_{\mathsf{u}}\big(\boldsymbol{\psi}(\mathbf{X}^n)\big)$ converges to $k$ almost surely. Note that $k \geq \frac{1}{\sum_{j=1}^{k} p_j^2}$, the almost-sure limit of $\Theta\big(\boldsymbol{\psi}(\mathbf{X}^n)\big)$, with equality iff $\mathbf{p} = U_k$. This follows from the Cauchy-Schwartz inequality: $1 = \left(\sum_{j=1}^{k} p_j\right)^2 \leq \left(\sum_{j=1}^{k} 1^2\right)\left(\sum_{j=1}^{k} p_j^2\right)$.

## V. A Pattern Classification Algorithm

Our proposed pattern classification algorithm is based on a slight relaxation of the necessary condition in Corollary 1.2. The algorithm is governed by a parameter $\epsilon > 0$, which can be used to control the trade-off between computational complexity and performance.

Recall from Section I that we call a pattern *uniform* if it has a PML distribution that is uniform; otherwise, we call the pattern *non-uniform*. Algorithm 2 below is our proposed pattern classification algorithm.

Clearly, if the input to the algorithm is a uniform pattern $\boldsymbol{\psi}$, then Corollary 1.2 shows that the algorithm correctly classifies the pattern as uniform. In this case, the distribution $U_k$ with $k = k^*_{\mathsf{u}}(\boldsymbol{\psi})$, which the algorithm would have to determine by a call to the procedure in Algorithm 1, is the true PML distribution for $\boldsymbol{\psi}$. If the input $\boldsymbol{\psi}$ is non-uniform, it is possible for the algorithm to classify it as uniform. In particular, if it happens that $\Theta(\boldsymbol{\psi}) < k^*_{\mathsf{u}}(\boldsymbol{\psi}) \leq \Theta(\boldsymbol{\psi}) + \epsilon$, so that $\boldsymbol{\psi}$ is guaranteed to be non-uniform, the CLASSIFY procedure still classifies it as uniform. The algorithm is designed to do this for two reasons:

**Algorithm 2**

**procedure** CLASSIFY($\boldsymbol{\psi}$)
  **if** $(k^*_{\mathsf{u}}(\boldsymbol{\psi}) \leq \Theta(\boldsymbol{\psi}) + \epsilon)$ **then**
    Classify $\boldsymbol{\psi}$ as Uniform
  **else**
    Classify $\boldsymbol{\psi}$ as Non-Uniform
  **end if**
**end procedure**

(a) For a pattern $\boldsymbol{\psi}$ classified as uniform, there is no further computational complexity involved in determining its presumed PML distribution $U_k$, with $k = k^*_{\mathsf{u}}(\boldsymbol{\psi})$.

(b) The distribution $U_k$, with $k = k^*_{\mathsf{u}}(\boldsymbol{\psi})$, is actually a good approximation for $\mathbf{p}^{\mathrm{PML}}(\boldsymbol{\psi})$, with the quality of the approximation getting better as $\epsilon \searrow 0$. We will justify this statement at the end of this section.

Thus, the parameter $\epsilon$ controls a trade-off between computational complexity of the classification algorithm and the approximation error obtained when a uniform distribution is used to approximate the PML distribution for a non-uniform pattern.

The next theorem provides some theoretical justification for our pattern classification algorithm.

**Theorem 3.** *Let $X_1, X_2, X_3, \cdots$ be a sequence of i.i.d. samples drawn from a distribution $\mathbf{p}$ with support $[k]$, i.e., $\mathbf{p} = (p_1, \ldots, p_k)$ with $p_j > 0$ for all $j \in [k]$.*

*(a) If $\mathbf{p} = U_k$, then the probability that Algorithm 2 declares the pattern $\boldsymbol{\psi}(\mathbf{X}^n)$ to be non-uniform goes to 0 as $n \to \infty$.*

*(b) If $\mathbf{p}$ is a non-uniform distribution, then for the parameter $\epsilon > 0$ chosen in Algorithm 2, the following statements hold:*

- *if $k - \frac{1}{\sum_{j=1}^{k} p_j^2} > \epsilon$, then the probability that the algorithm declares the pattern $\boldsymbol{\psi}(\mathbf{X}^n)$ to be uniform goes to 0 as $n \to \infty$;*
- *if $k - \frac{1}{\sum_{j=1}^{k} p_j^2} < \epsilon$, then the probability that the algorithm declares the pattern $\boldsymbol{\psi}(\mathbf{X}^n)$ to be uniform goes to 1 as $n \to \infty$;*

The proof follows in a routine manner from the fact that, almost surely, $\Theta\big(\boldsymbol{\psi}(\mathbf{X}^n)\big)$ converges to $\frac{1}{\sum_{j=1}^{k} p_j^2}$ and $k^*_{\mathsf{u}}\big(\boldsymbol{\psi}(\mathbf{X}^n)\big)$ converges to $k$. We omit the details.

The theorem above indicates that when the underlying

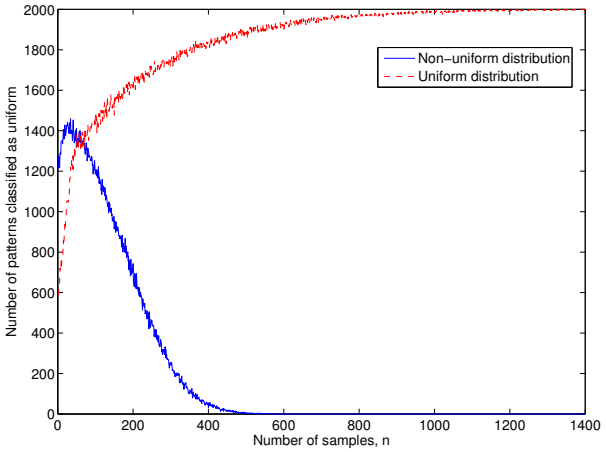


Fig. 1. A plot showing the performance of the pattern classifier for two underlying distributions with support size 500. The $y$-axis records the number, out of 2000 length-$n$ patterns generated independently from the underlying distribution, of patterns that were classified as having a uniform PML distribution.

distribution $\mathbf{p}$ is non-uniform and $k - \frac{1}{\sum_{j=1}^{k} p_j^2} < \epsilon$, then the classification algorithm will, with probability going to 1, misclassify a non-uniform pattern as uniform. In this case, the PML distribution of the pattern will be determined to be $U_k$, again with probability going to 1 (since $k_{\mathsf{u}}^*\big(\boldsymbol{\psi}(\mathbf{X}^n)\big)$ converges to $k$ almost surely). Now, consider the $\ell^2$ distance between $U_k$ and the underlying (non-uniform) distribution $\mathbf{p}$:

$$\|\mathbf{p} - U_k\|_2 = \sqrt{\sum_{j=1}^{k} \left(p_j - \frac{1}{k}\right)^2} = \sqrt{\sum_{j=1}^{k} p_j^2 - \frac{1}{k}}.$$

Since $k - \frac{1}{\sum_{j=1}^{k} p_j^2} < \epsilon$, we have

$$\sum_{j=1}^{k} p_j^2 - \frac{1}{k} < \frac{\epsilon \sum_{j=1}^{k} p_j^2}{k} \le \frac{\epsilon}{k}.$$

Hence, $\|\mathbf{p} - U_k\|_2 \le \sqrt{\epsilon/k}$. Thus, despite the misclassification of the pattern, the presumed PML distribution $U_k$ is a very good approximation (in terms of $\ell_2$ distance) of the underlying distribution $\mathbf{p}$.

## VI. Experimental Results

For experimental validation of our algorithm, we considered i.i.d. samples drawn from an underlying distribution $\mathbf{p}$. For each value of $n$ from 1 to 1400, we generated 2000 sets of $n$ independent samples drawn from $\mathbf{p}$. For each set of $n$ samples, we ran our classification algorithm on the length-$n$ pattern derived from the samples. We kept track of the number of times, out of 2000, that the algorithm classified the length-$n$ pattern as uniform. Figure 1 plots the results obtained for two different underlying distributions $\mathbf{p}$. One distribution is the uniform distribution on 500 symbols. The other distribution is a strongly non-uniform distribution supported on $\{1, 2, \ldots, 500\}$, given by $p_j = c/(50 + j)$ for $j = 1, 2, \ldots, 500$, where $c$ is a normalization constant required to make $\mathbf{p}$ a probability mass function. We used $\epsilon = 25$ in our classification algorithm; note that for the non-uniform distribution, $k - \frac{1}{\sum_{j=1}^{k} p_j^2} \approx 182.7$, which is much larger than $\epsilon$. The results clearly indicate that our pattern classification algorithm works well when the number of samples is sufficiently large.

## VII. A Conjecture

We would like to be so bold as to make the following conjecture, which is essentially a converse to Corollary 1.2. We admit that the conjecture is based on extremely limited experimental evidence.

**Conjecture 1.** *Suppose that the pattern $\boldsymbol{\psi}$ has a discrete PML distribution. If $k_u^*(\boldsymbol{\psi}) < \Theta(\boldsymbol{\psi})$, then the PML distribution of $\boldsymbol{\psi}$ is $U_k$, with $k = k_u^*(\boldsymbol{\psi})$.*

## References


[1] A. Orlitsky, N.P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1469–1481, July 2004.

[2] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang, "On estimating the probability multiset," draft manuscript, June 2011.

[3] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," in *Proc. 20th Conf. Uncertainty in Artificial Intelligence*, 2004.

[4] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang, "Limit results on pattern entropy," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 2954–2964, July 2006.

[5] J. Acharya, A. Orlitsky, and S. Pan, "Recent results on pattern maximum likelihood," *Proc. 2009 IEEE Inf. Theory Workshop (ITW'09)*, Volos, Greece, June 10–12, 2009, pp. 251–255.

[6] A. Orlitsky and S. Pan, "The maximum likelihood probability of skewed patterns," *Proc. 2009 IEEE Int. Symp. Inf. Theory (ISIT'09)*, Seoul, Korea, June 28 – July 3, 2009, pp. 1130–1134.

[7] J. Acharya, A. Orlitsky, and S. Pan, "The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns," *Proc. 2009 IEEE Int. Symp. Inf. Theory (ISIT'09)*, Seoul, Korea, June 28 – July 3, 2009, pp. 1135–1139.

[8] J. Acharya, H. Das, H. Mohimani, A. Orlitsky, and S. Pan, "Exact calculation of pattern probabilities," *Proc. 2010 IEEE Int. Symp. Inf. Theory (ISIT'10)*, Austin, Texas, USA, June 13–18, 2010, pp. 1498–1502.

[9] J. Acharya, H. Das, A. Orlitsky, and S. Pan, "Algebraic computation of pattern maximum likelihood," *Proc. 2011 IEEE Int. Symp. Inf. Theory (ISIT'11)*, Saint Petersburg, Russia, July 31 – Aug. 5, 2011, pp. 400–404.

[10] A. Orlitsky, Sajama, N.P. Santhanam, K. Viswanathan, and J. Zhang, "Algorithms for modeling distributions over large alphabets," *Proc. 2004 IEEE Int. Symp. Inf. Theory (ISIT'04)*, Chicago, USA, June 27 – July 2, 2004, p. 304.

[11] P.O. Vontobel, "The Bethe approximation of the pattern maximum likelihood distribution," *Proc. 2012 IEEE Int. Symp. Information Theory (ISIT'12)*, Cambridge, Mass., USA, July 1–6, 2012, pp. 2012–2016.

[12] P.O. Vontobel, "The Bethe permanent of a nonnegative matrix", *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1866–1901, March 2013.