The normalized maximum likelihood distribution of the multinomial model class with positive maximum likelihood parameters

Luosto, Panu

University of Helsinki, Department of Computer Science

2012


http://hdl.handle.net/10138/35373

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.
The normalized maximum likelihood distribution of the multinomial model class with positive maximum likelihood parameters

Panu Luostoa, Petri Kontkanena, Kerkko Luostob

aDepartment of Computer Science, University of Helsinki, Finland
bDepartment of Mathematics and Statistics, University of Helsinki, Finland

Abstract

The normalized maximum likelihood (NML) distribution is the distribution minimizing the worst-case redundancy of the code length compared to the unachievable optimal encoding. We give an effective way to calculate the NML distribution according to the multinomial model class in a configuration where the data is more restricted than in the usual setting. This situation, in which the maximum likelihood parameters of the multinomial distributions are always positive, is relevant e.g. in clustering applications. The corresponding NML distribution differs from the general multinomial NML distribution considerably when the number of parameters is large. Our main contribution is a recurrence relation with which this new NML can be calculated in \(O(n)\) time when the model class consists of Multinomial\((n, p_1, p_2, \ldots, p_k)\) distributions and the data belongs to the set \(\{(x_1, x_2, \ldots, x_k) \in \{1, 2, \ldots, n-k+1\}^k \mid \sum_{i=1}^k x_i = n\}\). We also prove that with a fixed \(n\) the parametric complexity of the model class is maximized with \(k = \lfloor n/4 \rfloor + 1\) or \(k = \lceil n/4 \rceil + 1\).

1. Introduction

A central problem in machine learning and statistics is model selection, which means selecting from a fixed set of models the model that fits a given data sample best. The minimum description length (MDL) principle [1, 2, 3, 4, 5] is a general framework for performing statistical inference. According to it, we should choose the model that allows us to encode both the data and the model with the shortest code.

In a simple encoding context, there is a set of distributions \(M = \{P(\cdot; \theta) \mid \theta \in \Theta\}\) with the domain \(X\), where \(\theta\) is a parameter vector. Normally no fixed element of \(M\) is a good choice for encoding, because the distribution dictating the code lengths has to be chosen before seeing the data. After having seen the data, we know that the solution leading to the shortest code length would have been the distribution with the maximum likelihood (ML) parameters. We call the ML solution optimal with regard to \(M\), but it can only be achieved with hindsight.

The normalized maximum likelihood (NML) distribution is the unique distribution minimizing the worst-case additional code length compared to the ML code length [4]. This excess code length is called the parametric complexity of \(M\). The NML distribution minimizes also the expected code length, when the expectation is taken with respect to
the worst possible data generating distribution \[6\]. However, it should be noted that sometimes the parametric complexity of a model class is infinite, and there is thus no corresponding NML distribution.

In this paper we discuss two different uses of the multinomial model class. In the first case there are no special restrictions for the data, especially it is possible that some of the ML parameters are zero. The corresponding NML distribution can be determined using a recurrence formula from \[7\]. \[8\] takes advantage of this distribution for determining the optimal cut point set for a histogram with variable width bins.

But there are also important applications in which the domain of the data is assumed to be such that the ML parameters are always positive. At the end of Section 3 we demonstrate that this kind of data configuration is natural for clustering with an unknown number of clusters. The main contribution of this paper is to derive an effective way to compute the NML for the multinomial model class in this case, and to show how the parametric complexity behaves as a function of the number of parameters. The results are equally applicable for sequences of length \(n\) that are modelled with \(n\) independent Multinomial(1, \(p_1\), ..., \(p_k\)) random variables, as well as for sequences of length \(k\) modelled with a Multinomial(\(n, p_1, \ldots, p_k\)) random variable.

This paper is structured as follows. In Section 2, we define the NML distribution according to the multinomial model class for the two configurations of data mentioned above. In Section 3, we give example applications for both versions of the multinomial NML. The subject of Section 4 is to derive a recurrence relation with which the normalizing sum of the multinomial model class with positive parameters can be calculated in \(O(n)\) time. In Section 5, we give the value of \(k\) with which the parametric complexity is maximized when \(n\) is fixed. We conclude the paper with a brief discussion in Section 6.

2. The NML of the multinomial model class

Consider a sequence of \(n\) repeated independent trials, each of which having the possible outcomes 1, 2, ..., \(k\) with the probabilities \(p_1, p_2, \ldots, p_k\), respectively, and hence \(\sum_{i=1}^{k} p_i = 1\). In other words, we have a distribution of sequences of \(n\) independent Multinomial(1, \(p_1\), ..., \(p_k\)) random variables. Let \(c = (c_1, c_2, \ldots, c_n) \in \{1, 2, \ldots, k\}^n\) be an outcome sequence, and let \(n_i(c) = |\{j : c_j \text{ is an element of } c, c_j = i\}|\) be the number of occurrences of \(i\) in \(c\). Writing \(p = (p_1, p_2, \ldots, p_k)\), the probability of \(c\) is

\[ P_k(c; p) = \prod_{i=1}^{k} p_i^{n_i(c)}. \]

We use here and throughout this article the convention \(0^0 \equiv 1\).

Assume now that the probabilities \(p_1, p_2, \ldots, p_k\) are unknown. The model class in the question is the set

\[ M_k = \{P_k(\cdot; p) \mid p \in [0, 1]^k, \sum_{i=1}^{k} p_i = 1\}. \quad (1) \]

It is a simple multidimensional optimization problem to show that the parameter vector which maximizes the likelihood with respect to \(c\) is \(\hat{p}(c) = (n_1(c), n_2(c), \ldots, n_k(c))/n\).
As stated in the previous section, the worst-case-optimal way to encode $c$ according to $M_k$ is the normalized maximum likelihood distribution, defined as

$$P^\text{NML}_k(c) := \frac{P_k(c; \hat{p}(c))}{C_0(k, n)} \quad (2)$$

where

$$C_0(k, n) = \sum_{c \in \{1, \ldots, k\}^n} P_k(c; \hat{p}(c))$$

$$= \sum_{m_1 + \cdots + m_k = n, \atop m_1, \ldots, m_k \geq 1} \frac{n!}{m_1! \cdots m_k!} \prod_{i=1}^k \left( \frac{m_i}{n} \right)^{m_i} \quad (3)$$

One sees immediately that $\log C_0(k, n)$ is also the parametric complexity of the model class with Multinomial($n, p_1, \ldots, p_k$) distributions. Values of the function $C_0$ can be computed effectively using the recursion formula $C_0(k+2, n) = C_0(k+1, n) + (n/k) C_0(k, n)$ [7].

However, there are situations in which the NML defined in (2) is not the most appropriate choice. In some applications, it is natural to assume that the data belongs to the set $\{c \in \{1, 2, \ldots, k\}^n \mid n_i(c) \geq 1 \text{ for all } i \in \{1, 2, \ldots, k\} \}$, and thus the maximum likelihood parameter vector $\hat{p}(c)$ never contains zero elements. In that case we must normalize $P_k(c; \hat{p}(c))$ with the sum

$$C_1(k, n) = \sum_{m_1 + \cdots + m_k = n, \atop m_1, \ldots, m_k \geq 1} \frac{n!}{m_1! \cdots m_k!} \prod_{i=1}^k \left( \frac{m_i}{n} \right)^{m_i}$$

$$= \frac{n!}{n^n} \sum_{m_1 + \cdots + m_k = n, \atop m_1, \ldots, m_k \geq 1} \prod_{i=1}^k \frac{m_i^{m_i}}{m_i!}, \quad (4)$$

and the corresponding NML is

$$\tilde{P}^\text{NML}_k(c) := \frac{P_k(c; \hat{p}(c))}{C_1(k, n)} \quad (5)$$

In the next section, we give as examples two applications that illustrate the use of $P^\text{NML}_k$ and $\tilde{P}^\text{NML}_k$.

### 3. Example applications of the two variants of the multinomial NML distribution

Our two example applications can be considered as clustering problems in which the number of clusters $k \in \{1, 2, \ldots, k_0\}$ is unknown.

#### 3.1. Choosing the number of bins for a regular histogram

Our first example studies choosing the number of bins for a regular histogram, in which the bins are equally wide. An interested reader may refer to [8] for information
about irregular MDL histograms. Assume that the range of the data is \([a, b] \subset \mathbb{R}\), and let the data sequence of independent and identically distributed points be \(x = (x_1, x_2, \ldots, x_n) \in [a, b]^n\). In a histogram with \(k\) bins, the bins correspond to the intervals \([a, a + w_k[, [a + w_k, a + 2w_k[, \ldots, [a + (k - 1)w_k, b]\) where \(w_k = (b - a)/k\). Let \(c_k(x) = (c_{k1}, c_{k2}, \ldots, c_{kn}) \in \{1, 2, \ldots, k\}^n\) indicate the assignments of the elements of \(x\) into the bins, so that \(c_{ki} = j\) if \(x_i \in [a + (j - 1)w_k, a + jw_k]\). In the histogram model, on the condition that \(k\) and \(c_{ki}\) are known, the density of the point \(x_i\) is uniform within the interval corresponding to the bin \(c_{ki}\).

Let \(k_0 \in \{1, 2, \ldots, \}\), and let \(P_{k}^{\text{NML}}\) be defined as in (2). The model class selection problem is to choose the number of bins \(k \in \{1, 2, \ldots, k_0\}\). As a criterion we use maximization of the density

\[
f(x, k) = \frac{1}{k_0} P_{k}^{\text{NML}}(c_k(x)) \left( \frac{1}{w_k} \right)^n
\]  

with respect to \(k\). Equivalently, we choose the \(k\) that minimizes \(-\log f(x, k)\), which is the analogue for code lengths in the field of continuous distributions. In (6) we encode \(k\) according to the uniform distribution in \(\{1, 2, \ldots, k_0\}\). Note also that with \(P_{k}^{\text{NML}}\) we can encode bin assignment sequences in which \(n_i(c_k(x)) = 0\) for some \(i \in \{1, 2, \ldots, k\}\). Requiring that always \(n_i(c_k(x)) > 0\) would clearly not be reasonable in this context. As an example one may think of a sequence \(x\) half of the elements of which belong to the interval \([a, a + \epsilon]\) and the rest to the interval \([b - \epsilon, b]\), where \(\epsilon > 0\) is a small constant.

### 3.2. A probabilistic quality measure for a clustering

In our second example, we derive a probabilistic criterion for choosing the best clustering in a more general clustering context than in Subsection 3.1. Let \(X = (x_1, x_2, \ldots, x_n) \in (\mathbb{R}^d)^n\) be a data sequence, and let \(k\) be the number of clusters. We assume for simplicity that the cluster distributions are independent. For every \(j \in \{1, 2, \ldots, n\}\) there is a single density \(f_j : (\mathbb{R}^d) \rightarrow [0, \infty]\) that is used for encoding a cluster data sequence of length \(j\). For example, \(f_j\) can be the NML of some model class. Let \(c_k = (c_{k1}, c_{k2}, \ldots, c_{kn}) \in \{1, 2, \ldots, k\}^n\) be an integer sequence that indicates the cluster assignments. We write the subsequence of \(X\) corresponding to cluster \(j\) as \(X_j(c_k)\). So the element \(x_i\) is in the sequence \(X_j(c_k)\) if and only if \(c_{ki} = j\).

Assume that the elements of \(c_k\) carry no information of the data itself, unlike in the histogram example of the previous subsection. Then, the actual values of the elements in \(c_k\) are not important, e.g., \((3, 1, 1, 2)\) and \((2, 3, 3, 1)\) refer to the same clustering of four data points into three clusters. We encode thus not the sequence \(c_k\) but the equivalence class \([c_k] = \{c \in \{1, 2, \ldots, k\}^n \mid c\) represents the same clustering as \(c_k\}\). Given \(k\), we can also assume that there are no empty clusters, that is, \(n_j(c_k) \geq 1\) for all \(j \in \{1, 2, \ldots, k\}\). That allows us to define the conditional probability of \([c_k]\) according to the multinomial model class as

\[
P([c_k] \mid K = k) = k! \tilde{P}_k^{\text{NML}}(c_k)
\]  

where \(\tilde{P}_k^{\text{NML}}\) is defined as in (5).
Finally, our criterion for finding the best number of clusters $k \in \{1, 2, \ldots, k_0\}$ and the best clustering $[c_k]$ for $X$ is to maximize with respect to $(k, [c_k])$ the density

$$g(X, k, [c_k]) = \frac{1}{k_0} k! \tilde{P}^{NML}_k (c_k) \prod_{j=1}^{k} f_{n_j}(X_j(c_k)) \quad (8)$$

where $n_j = n_j(c_k)$ is the length of the cluster subsequence $X_j(c_k)$.

A similar density to (8) for a clustering was introduced in [9], where $P^{NML}_k$ was used instead of $\tilde{P}^{NML}_k$. But because it is assumed in [9] that none of the clusters is empty, the new density (8) is a more natural choice.

4. A recurrence relation for the normalizing sum

In this section, we derive a recursion relation for the sum $C_1(k, n)$, which is defined in (4). We use the mathematical technique of generating functions [10] in a similar fashion as in [7].

**Theorem 1.** Let $n \in \{3, 4, \ldots, \}$. For all $k \in \{1, 2, \ldots, n-2\}$ it holds

$$C_1(k+2, n) + 2C_1(k+1, n) = \left(\frac{n}{k} - 1\right) C_1(k, n).$$

**Proof.** Let $k \in \{1, 2, \ldots, \}$. Let the generating function for the sequence $(n^n/n!)_{n=0}^{\infty}$ be

$$B(z) = \sum_{n=0}^{\infty} \frac{n^n}{n!} z^n.$$

Let

$$T(z) = \sum_{n=1}^{\infty} \frac{n^{n-1}}{n!} z^n \quad (9)$$

be the so called tree function [11], which is closely related to Lambert W function [12]. It can be shown with the Lagrange inversion formula that $T(z) = z e^{T(z)} [10]$. Thus, $T'(z) = e^{T(z)} + T(z)T'(z)$ and $z T'(z) = T(z)/(1 - T(z))$. On the other hand, we see directly from (9) that $z T'(z) = B(z) - 1$. It follows that $B(z) - 1 = T(z)/(1 - T(z))$. Now,

$$\left(\frac{T(z)}{1 - T(z)}\right)^k$$

$$= (B(z) - 1)^k$$

$$= \prod_{i=1}^{k} \sum_{m_i=1}^{\infty} \frac{m_1^{m_1}}{m_1!} z^{m_1}$$

$$= \sum_{n=k}^{\infty} \sum_{m_1, \ldots, m_k \geq 1} \frac{m_1^{m_1} \ldots m_k^{m_k}}{m_1! \ldots m_k!} z^n$$

$$= \sum_{n=k}^{\infty} S(k, n) z^n \quad (11)$$

5
where \( S(k, n) = (n^n/n!)C_1(k, n) \). Next we differentiate both sides of the equation (10) = (11) with respect to \( z \) and multiply the result by \( z \). The left-hand side will be

\[
\frac{z}{dz} \left( \frac{T(z)}{1 - T(z)} \right)^k \]

\[
= z k \left( \frac{T(z)}{1 - T(z)} \right)^{k-1} \frac{T'(z)(1 - T(z)) + T(z)T''(z)}{(1 - T(z))^2}
\]

\[
= k \frac{T(z)^{k-1}}{(1 - T(z))^{k+1}} z T'(z) \]

\[
= k \left( \frac{T(z)}{1 - T(z)} \right)^{k+2} + 2 \left( \frac{T(z)}{1 - T(z)} \right)^{k+1} + \left( \frac{T(z)}{1 - T(z)} \right)^k,
\]

and the right-hand side \( \sum_{n=k}^{\infty} nS(k, n)z^n \). Combining these, and using (10) = (11) in the left-hand side terms we obtain

\[
k \sum_{n=k+2}^{\infty} S(k + 2, n)z^n + 2k \sum_{n=k+1}^{\infty} S(k + 1, n)z^n + k \sum_{n=k}^{\infty} S(k, n)z^n = \sum_{n=k}^{\infty} nS(k, n)z^n.
\]

From (12) we get for all \( n \in \{k+2, k+3, \ldots\} \) that \( k \sum_{n=k+2}^{\infty} S(k + 2, n)z^n + 2k \sum_{n=k+1}^{\infty} S(k + 1, n)z^n + k \sum_{n=k}^{\infty} S(k, n)z^n = nS(k, n) \). Theorem 1 follows trivially.

In practice, there are two possibilities to use the recursion formula, proceeding either in ascending or descending order. In the first case, we start with \( C_1(1, n) = 1 \) and \( C_1(2, n) \). The sum \( C_1(2, n) \) can be computed according to (4) in \( O(n) \) time if we assume that exponentiation is a constant time operation. Alternatively, we can start with \( C_1(n, 1) = n!/n^n \) and \( C_1(n - 1, 1) = 2(n - 1)C_1(n, n) \), which seems to be more stable numerically. If \( n \) is large, the factor \( n!/n^n \) can be approximated with Stirling’s formula.

5. Maximum point of \( C_1(k, n) \) with a fixed \( n \)

It is easy to see from the definition in (3) that \( C_0(k, n) \) grows monotonically as a function of \( k \) when \( n \) is fixed. In contrast, the sum \( C_1(k, n) \) behaves quite differently. We conjecture that \( C_1(k, n) \) is maximized with a fixed \( n \) when \( k = \lfloor n/4 \rfloor + 1 \). It is easy to verify that the conjecture is true for \( n \in \{1, 2, 3\} \). The following theorem proves that the conjecture holds also when \( n \) is divisible by 4. Interestingly, Theorem 2 holds for any positive function \( c : \{1, 2, \ldots, n\} \to \mathbb{R}_+ \) that satisfies the recurrence relation of the same form as in Theorem 1.

Let \( n \in \{4, 5, \ldots\} \), and let \( c : \{1, 2, \ldots, n\} \to \mathbb{R}_+ = [0, \infty[ \) be a function satisfying the recurrence relation

\[
c(k + 2) + 2c(k + 1) = \left( \frac{n}{k} - 1 \right) c(k)
\]

for all \( k \in \{1, 2, \ldots, n - 2\} \).
Theorem 2. Function $c$ gets its maximum value when $k = \lfloor n/4 \rfloor + 1$ or $k = \lceil n/4 \rceil + 1$.

Proof. We first fix some notation. Consider the ratio $u : \{1, 2, \ldots, n-1\} \to \mathbb{R}_+$, $u(k) = c(k)/c(k+1)$, the coefficient $\alpha : \{1, 2, \ldots, n-2\} \to \mathbb{R}_+$, $\alpha(k) = (n/k - 1)^{-1}$, and let $x : \{1, 2, \ldots, n-2\} \to \mathbb{R}_+$,

$$x(k) = \alpha(k) + \sqrt{\alpha(k) + \alpha(k)^2}.$$  

Also, for $k \in \{1, 2, \ldots, n-2\}$ let $f_k : \mathbb{R}_+ \to \mathbb{R}_+$,

$$f_k(t) = \alpha(k)(t^{-1} + 2).$$

It is easy to verify that $\alpha$ and $x$ are strictly increasing functions, and that for all $k \in \{1, \ldots, n-2\}$ the function $f_k$ is strictly decreasing. Using (13) we see that for $k \in \{1, 2, \ldots, n-2\}$ the function $u$ satisfies the recurrence relation

$$u(k) = \alpha(k) \frac{\alpha(k)^{-1} c(k)}{c(k+1)}$$
$$= \alpha(k) \frac{c(k+2) + 2c(k+1)}{c(k+1)}$$
$$= \alpha(k)(u(k+1)^{-1} + 2)$$
$$= f_k(u(k+1)). \quad (14)$$

Note that $x(k)$ is a fixed point of $f_k$, in other words,

$$f_k(x(k)) = x(k). \quad (15)$$

The next lemma includes a rough estimate of the ratio $u(k)$.

Lemma 1. For $k \in \{1, 2, \ldots, n-3\}$, it holds that $\{x(k), u(k)\} \subset [2\alpha(k), 2\alpha(k) + 1/2]$. Furthermore, $u(n-2) > 2\alpha(n-2)$.

Proof. For $k \in \{1, 2, \ldots, n-2\}$,

$$2\alpha(k) < \alpha(k) + \sqrt{\alpha(k) + \alpha(k)^2}$$
$$= x(k)$$
$$< \alpha(k) + \sqrt{(1/2 + \alpha(k))^2}$$
$$= 2\alpha(k) + 1/2.$$  

Equation (14) and the positivity of $u$ yield $u(k) = \alpha(k)(u(k+1)^{-1} + 2) > 2\alpha(k)$. And thus when $k \leq n-3$, it holds

$$u(k) = \alpha(k)(u(k+1)^{-1} + 2)$$
$$< \alpha(k)((2\alpha(k) + 1)^{-1} + 2)$$
$$< \alpha(k)((2\alpha(k))^{-1} + 2)$$
$$= 2\alpha(k) + 1/2.$$  

In the last inequality we used the fact that $\alpha$ is strictly increasing. □
Lemma 2. For \( k \in \{2, 3, \ldots, n-3\} \), the inequality \( x(k-1) \leq f_k(x(k+1)) \) holds.

Proof. A simple calculation yields \( f_k(x(k+1)) = \alpha(k) + \alpha(k) \sqrt{\frac{1}{\alpha(k+1)}} + 1 \). Because for all \( j \in \{1, 2, \ldots, n-4\} \) it holds \( \alpha(j) + \alpha(j)^2 = nj/(n-j)^2 \), we get for \( k \in \{2, 3, \ldots, n-3\} \) that

\[
f_k(x(k+1)) = \alpha(k) + \alpha(k) \sqrt{\frac{1}{\alpha(k+1)}} + 1
= \alpha(k) + \sqrt{\frac{k^2n}{(n-k)^2(k+1)}}
\geq \alpha(k-1) + \sqrt{\frac{(k^2-1)n}{(n-(k-1))^2(k+1)}}
= \alpha(k-1) + \sqrt{\alpha(k-1) + \alpha(k-1)^2}
= x(k-1).
\]

The next lemma is the key for our estimation.

Lemma 3. Let \( n \geq 5 \). Then for all \( k \in \{1, 2, \ldots, n-4\} \) it holds that \( x(k) < u(k+1) < x(k+1) \). Additionally, \( x(n-3) < u(n-2) \).

Proof. We prove the statement by backward induction with regard to \( k \). The base of the induction is \( k = n-4 \). Assuming \( n \geq 5 \), we can estimate that

\[
2\alpha(n-4) + \frac{1}{2} = \frac{n}{2} - \frac{3}{2} \leq \frac{n}{2} - \frac{3}{2} + \frac{1}{6} (n-5)
< \frac{3}{2} n - 2
= 2\alpha(n-3).
\]

Using Lemma 1, we obtain

\[
x(n-4) < 2\alpha(n-4) + \frac{1}{2}
< 2\alpha(n-3)
< u(n-3).
\]

For the upper limit of \( u(n-3) \), we start by estimating that

\[
2\alpha(n-3) + \frac{1}{2} = \frac{2}{3} \cdot \frac{n}{2} - \frac{3}{2} \leq \frac{2}{3} \cdot \frac{n}{2} - \frac{3}{2} + \frac{1}{3} (n-5)
< n-2
= 2\alpha(n-2).
\]
which yields
\[ x(n - 3) < 2\alpha(n - 3) + \frac{1}{2} \]
\[ < 2\alpha(n - 2) \]
\[ < u(n - 2). \]

Because \( f_{n-3} \) is a strictly decreasing function, \( x(n - 3) < u(n - 2) \) implies \( f_{n-3}(u(n - 2)) < f_{n-3}(x(n - 3)) \), or equivalently, \( u(n - 3) < x(n - 3) \) on the basis of (14) and (15). We saw that the statement was true for \( k = n - 4 \), and we proved the inequality \( x(n - 3) < u(n - 2) \).

Assume that Lemma 3 holds for \( k > 1 \). Thus \( x(k) < u(k + 1) < x(k + 1) \), and because \( f_k \) is strictly decreasing, \( f_k(x(k + 1)) < f_k(u(k + 1)) < f_k(x(k)) \). Using Lemma 2, (14) and (15), we get \( x(k - 1) \leq f_k(x(k + 1)) < f_k(u(k + 1)) = u(k) < f_k(x(k)) = x(k) \). Hence, the statement holds for the value \( k - 1 \). □

**Lemma 4.** a) If \( k \in \mathbb{Z}_+ \), \( k \leq n/4 \), then \( u(k) < 1 \). b) If \( k \in \mathbb{Z}_+ \), \( n/4 + 1 \leq k \leq n - 2 \), then \( u(k) > 1 \).

**Proof.** If \( n = 4 \), we have to prove that \( u(1) < 1 \) and \( u(2) > 1 \). Lemma 1 yields \( u(2) > 2\alpha(2) = 2 \). Then with (14) we get \( u(1) = \alpha(1)(u(2)^{-1} + 2) < 1/3 \cdot (1/2 + 2) = 5/6 \).

Assume from now on that \( n \geq 5 \).

a) Let \( k \in \mathbb{Z}_+ \), \( k \leq n/4 \). If \( k = 1 \), it follows from Lemma 1 that
\[ u(1) < 2\alpha(1) + \frac{1}{2} \]
\[ = \frac{2}{n - 1} + \frac{1}{2} \]
\[ \leq \frac{2}{5 - 1} + \frac{1}{2} \]
\[ = 1. \]

Otherwise, \( 1 \leq k - 1 \leq n/4 - 1 + (3/4)(n - 5) < n - 4 \). Lemma 3 can thus be used, and
\[ u(k) < x(k) \]
\[ = \alpha(k) + \sqrt{\alpha(k) + \alpha(k)^2} \]
\[ \leq 1/3 + \sqrt{1/3 + (1/3)^2} \]
\[ = 1 \]

because \( \alpha(k) = (n/k - 1)^{-1} \leq (4 - 1)^{-1} = 1/3 \).

b) Let \( k \in \mathbb{Z}_+ \), \( n/4 + 1 \leq k \leq n - 2 \). Using Lemma 3 we get
\[ u(k) > x(k - 1) \]
\[ = \alpha(k - 1) + \sqrt{\alpha(k - 1) + \alpha(k - 1)^2} \]
\[ \geq 1/3 + \sqrt{1/3 + (1/3)^2} \]
\[ = 1. \]

because \( \alpha(k - 1) = (n/(k - 1) - 1)^{-1} \geq (4 - 1)^{-1} = 1/3 \). □
Proof of Theorem 2. According to Lemma 4 and the definition of u, function c is strictly increasing in the set \{1, 2, \ldots, \lfloor n/4 \rfloor + 1\}. On the other hand, function c is strictly decreasing in the set \{\lceil n/4 \rceil + 1, \ldots, n-1\}. Recalling that \( n \geq 4 \), and using the recurrence relation in (13) yields
\[
c(n) < c(n) + 2c(n-1) = \frac{2}{n-2} c(n-2) \leq \frac{2}{4-2} c(n-2) = c(n-2).
\]
Thus, the maximum value of u in the set \{\lceil n/4 \rceil + 1, \lfloor n/4 \rfloor + 2, \ldots, n\} is \( c(\lceil n/4 \rceil + 1) \). Because \{1, 2, \ldots, \lfloor n/4 \rfloor + 1\} \cup \{\lceil n/4 \rceil + 1, \ldots, n\} = \{1, 2, \ldots, n\}, the theorem is proved. \( \square \)

6. Discussion

Figure 1 illustrates the parametric complexities \( \log_2 C_0 \) and \( \log_2 C_1 \) as functions of \( k \) when \( n = 100 \). In the clustering context of Subsection 3.2, the extremes with one cluster or \( n \) clusters are both special cases. For example, we get from (7) that \( P([1, 2, \ldots, n] | K = n) = n!(1/n)^n/(n!/n^n) = 1 \), as one would expect. But using \( C_0(n, n) \) for normalizing the maximum likelihood in this case would yield a very long code length. Using \( P_k^{NML} \)
for encoding where $\hat{P}_N^{\text{NML}}$ would be a more natural choice may not have major practical consequences when $k$ is very small compared to $n$, but in applications in which larger values of $k$ are viable choices, the situation is likely to be different.

References