UNIVERSITY OF CALIFORNIA, SAN DIEGO

On The Theory and Application of Pattern Maximum Likelihood

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

in

Computer Science

by

Shengjun Pan

Committee in charge:

Professor Alon Orlitsky, Chair
Professor Ian Abramson
Professor Fan Chung Graham
Professor Ron Graham
Professor Russell Impagliazzo

2012

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

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2012

EPIGRAPH

书 山 有 路 勤 为 径
学 海 无 涯 苦 作 舟

— 韩愈

Live and learn.

— Han Yu

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First I would like to thank all my doctoral committee for sparing time out of their busy schedules for both my qualifying exam and final defense. In addition, thanks to Professor Ian Abramson for his interest in the research subject and patience during our individual presentation for my qualifying exam. Thanks to Professor Fan Chung Graham for introducing the best English tutor to me, and for her general suggestions on research and career. Thanks to Professor Ron Graham for taking me as a TA for his interesting and fun courses. Thanks to Professor Russell Impagliazzo for letting me learn a lot from his course on computational complexity.

I would like to express my special thanks to my advisor Professor Alon Orlitsky. I have learned many things from him which have benefited me during my study for Ph.D., and I believe they will continue to benefit me in my future career and life. The first thing I learned from him, when I was working on the skewed patterns, is that *difficulty is not equal to impossibility*. He has helped me improve my research skills, not just for patterns but as tools to learn and research new things. I would like also thank him for his word-by-word guidance on technical writing. There are more, but due to limited space I will leave out many other things I want to thank him for.

I would like to thank all my current labmates Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, and Ananda Theertha Suresh for their help and collaboration on various research topics. Thanks to them for their brilliant ideas and skills and the fun time in the last five/six years. I would like also thank the previous labmates Anand Dhulipala, Nikola Jevtic, Sajama, Prasad Santhanam, Krishnamurthy Viswanathan, and Junan Zhang for their kind help and communication even though they are no long in the lab.

At last I would like to thank my parents for their understanding and support of my oversea study. They always believe that I should get a better education to have a better future. Despite of their financial burdens they pushed forward my education, in contrast to some other parents in the same village who had to let their children drop out of school. For that they have my life-time gratitude.

Chapter 4 appeared, or will appear, partially in (i) The maximum likelihood probability of skewed patterns, Alon Orlitsky and Shengjun Pan, *IEEE International Symposium on Information Theory*, 2009. (ii) Recent results on pattern maximum likelihood, Jayadev Acharya, Alon Orlitsky and Shengjun Pan, *IEEE Information Theory Worshop, Volos*, Greece, 2009. (iii) Exact calculation of pattern probabilities, Jayadev Acharya, Hirakendu Das, Hosein Mohimani, Alon Orlitsky and Shengjun Pan, *IEEE International Symposium on Information Theory*, 2010. (iv) On the number of even and odd connected graphs, Philippe Flajolet, Alon Orlitsky and Shengjun Pan, *In preparation*, 2012.

Chapter 6 appeared partially in Pattern maximum likelihood: computation and experiments, Alon Orlitsky, Shengjun Pan, Sajama, Narayana Prasad Santhanam, Krishnamurthy Viswanathan and Junan Zhang, *In preparation*, 2012.

Chapter 7 appeared partially in Estimating multiset of Bernoulli processes, Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky and Shengjun Pan, *Submitted to ISIT*, 2012.

VITA

2002	B.S. in Mathematics, Peking University, China
2006	M.Math in Combinatorics and Optimization, University of Waterloo
2012	Ph.D. in Computer Science, University of California, San Diego

SELECTED PUBLICATIONS

Competitive closeness testing, J. Acharya, H. Das, A. Jafarpour, A. Orlitsky and S. Pan, *Conference On Learning Theroy*, 2011.

Algebraic calculation of maximum likelihood of small patterns, J. Acharya, H. Das, A. Orlitsky and S. Pan, *IEEE International Symposium on Information Theory*, 2011.

Classification using pattern probability estimators, J. Acharya, H. Das, A. Orlitsky, S. Pan and N. P. Santhanam, *IEEE International Symposium on Information Theory*, 2010.

Exact calculation of pattern probabilities, J. Acharya, H. Das, G. H. Mohimani, A. Orlitsky and S. Pan, *IEEE International Symposium on Information Theory*, 2010.

On reconstructing a string from its substring compositions, J. Acharya, H. Das, O. Milenkovic, A. Orlitsky and S. Pan, *IEEE International Symposium on Information Theory*, 2010.

The maximum likelihood probability of skewed patterns, A. Orlitsky and S. Pan, *IEEE International Symposium on Information Theory*, 2009.

The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns, J. Acharya, A. Orlitsky and S. Pan, *IEEE International Symposium on Information Theory*, 2009.

Recent results on pattern maximum likelihood, J. Acharya, A. Orlitsky and S. Pan, *IEEE Information Theory Worshop, Volos*, Greece, 2009.

The crossing number of K_{11} is 100, S. Pan and R. B. Richter, *J. Graph Theory*, 56(2):128-134, 2007.

The convex hull of every optimal pseudolinear drawing of K_n is a triangle, J. Balogh, J. Leaños, S. Pan, R. B. Richter, and G. Salazar, Australas. *J. Combin.*, 38:155-162, 2007.

ABSTRACT OF THE DISSERTATION

On The Theory and Application of Pattern Maximum Likelihood

by

Shengjun Pan

Doctor of Philosophy in Computer Science

University of California, San Diego, 2012

Professor Alon Orlitsky, Chair

Pattern Maximum Likelihood (PML) is a method of probability estimation that works well for large alphabets. It does not assume that all elements from the unknown alphabet have been observed. PML outperforms the traditional Maximum Likelihood for sequences, and it is particularly useful when the sample size is small.

In this dissertation we study both the theory and application of PML. For the theory part, we extend the the previous results on the properties of PML, and also show how to find the PML distributions analytically for patterns of simple forms. For general patterns, PML probabilities can be approximated using a previously developed EM algorithm, which we will prove to be equivalent to

a generalized Gradient Ascend Method. We also use the algorithm to conduct experiments on different distributions and evaluate the performance of PML.

In addition, we investigate the calculation of pattern probability. We show that the pattern probability is closely related to symmetric polynomials, and it can be written as a summation over graphs using power sums. Along the way we reveal a relation between pattern probability and the enumeration of certain connected graphs as well as inversion-free trees.

For applications, we show how PML can be used to predict the number of new symbols that would appear in a future sample. We conduct experiments on various distributions and compare PML to the method of Good & Toulmin and the method of Efron & Thisted. We demonstrate that PML outperforms the other methods even if the future sample size is large. Finally we apply PML to authenticating the authorship of the Taylor poem, attributed to Shakespeare, and conclude that it is consistent with Efron and Thisted's models.

PML deals with samples from a single distribution. In the last part of this dissertation we extend PML to set-patterns where multiple samples are observed from concurrent Bernoulli processes. Analogous to the single-process patterns, we show that for certain forms of set-patterns we can find the exact Set-pattern Maximum Likelihood (SPML) probabilities analytically. Furthermore, for general set-patterns we extend the previous EM algorithm to approximate the SPML probabilities. We also show that for samples taken from Poisson distributions the set-pattern is reduced to the single-process pattern problem.

Chapter 1

Introduction

Predicting the future has been of interest for a long time in human history. Probability estimation is the scientific approach of *guessing*, statistically, the underlying distribution based on observations in the past, and use it to estimate the likelihood of future events. It has applications in a variety of fields, such as weather forecast, economics, finance, etc. Particularly, in the age of computers it has been widely used in data compression, machine learning, communication, Internet, and other areas. For example, in data compression probability estimation allows us to assign shorter codewords to more frequent symbols so as to reduce the size of the compressed data. Another example is online advertising where advertisers would like to maximize the probability that an Internet user clicks through their ads.

A probability distribution can be regarded as two parts: the multiset of probabilities, and the association between the probabilities and the symbols in the underlying alphabet. Many applications require only the probability multiset. For example, a biologist might want to estimate the number of endangered species in an area, a bank may be interested in determining the proportion of customers with potentially high investments, and a shopping mall could benefit from estimating the amount of customers before increasing stocks. In all these applications, it is sufficient to estimate the probabilities as a multiset; the probability of a specific object is not of concern.

Sequence Maximum Likelihood (SML) is a commonly used method for estimating probabilities. It finds the distribution that maximizes the probability of

the observed sequence. SML is useful when all symbols appear sufficiently many times. However, there are unavoidable drawbacks when the sample is taken from a large data set where many symbols have low frequencies. Furthermore, SML always assigns zero probabilities to unseen symbols.

More practical approaches have been proposed to work better with samples from large alphabets. One line of work started from Fisher [FCW43], followed by Good and Toulmin [GT56], and Efron and Thisted [ET76, TE87]. A comprehensive survey can be found in Bunge and Fitzpatrick [BF93]. A more recent approach by Valiants appeared in [Val08, VV11a, VV11b].

An information-theoretically motivated method was pursued by Orlitsky *et al.* in [OSZ04, OSVZ04] with more recent development in [OP09, AOP09, ADJ$^+$11, OSVZ12, OPS$^+$12]. Their approach was based on the observation that, since we do not care about the association between the probabilities and the underlying symbols, the information that matters is how different symbols repeat. For example, intuitively the estimate of probabilities for the *i.i.d.* sequences *aba* should not differ from that for the sequence $@ \wedge @$. Thus we can *extract* the crucial information by replacing the symbols with their order of appearance, called the *pattern*. For example, The pattern for both *aba* and $@ \wedge @$ are 121. It is a way of saying that these two sequences repeat alike. Analogous to SML, which maximizes the sequence probability, the *Pattern Maximum Likelihood* maximizes the pattern probability.

To see the advantage of PML intuitively, consider estimating the distribution of human DNAs. If we take a sample of 100 DNAs from the population, with high probability we would get 100 distinct DNAs (without twins). Then SML would say that each observed DNA has probability 1%, and any DNA not observed has probability 0. This is clearly far from the truth since the pool of DNAs is so vast that a sample of size 100 is nowhere close to being sufficient. On the other hand, assuming zero-knowledge of the underlying alphabet, PML would say that the sample has a pattern that *all symbols are distinct*, and such a pattern would most likely rise from a continuous distribution. Indeed, if we sample from a continuous distribution over real numbers, with probability 1 we wouldn't see any

repetitions, which is exactly the case as in sampling DNAs.

Note that SML always assign zero probabilities to symbols that have not been observed, and thus it tends to overestimate the probabilities of observed symbols. As a comparison, PML does not assume that the underlying support size is known; it maximizes the pattern probability over all discrete distributions, with a possible continuous part, of all support sizes. Thus PML can be potentially useful for estimating the support size of the underlying distribution, as well as missing probabilities.

Let's take a look at another example in text classification. Suppose we have observed two sequences *aaabbb* and $\alpha\alpha\beta\beta\gamma\gamma$ from two distributions P_1 and P_2 respectively. We would like to classify a third sequence $xxyyzz$ as from P_1 or P_2. Using SML we would obtain two estimates $P^1_{\text{SML}} = \left(a \to \frac{1}{2}, b \to \frac{1}{2}\right)$ and $P^2_{\text{SML}} = \left(\alpha \to \frac{1}{3}, \beta \to \frac{1}{3}, \gamma \to \frac{1}{3}\right)$. Since the sequences have no common observations, both $P^1_{\text{SML}}(xxyyzz)$ and $P^2_{\text{SML}}(xxyyzz)$ are zero, thus SML is unable to classify $xxyyzz$. On the other hand, PML would classify $xxyyzz$ as from P_2, which seems more reasonable, since it has the same pattern as $\alpha\alpha\beta\beta\gamma\gamma$, namely 112233.

The results are deployed in the remaining chapters as follows.

- In Chapter 2 we formally define pattern and PML, introduce necessary notions, and describe previous known results.
- In Chapter 3, we extend the previous results on the PML support size and the continuous probability.
- In Chapter 4, we find the PML for patterns of simple forms.
- In Chapter 5 we analyze the computational complexity of the deterministic calculation of pattern probabilities.
- In Chapter 6 we describe the EM algorithm and Metropolis algorithm used to approximate PML distributions for general patterns, and show how to apply PML to practical problems.
- Finally, in Chapter 7 we extend techniques and results for patterns to set-patterns where the observation comes from multiple sampling processes.
- Some of the technical proofs are put in Appendix A.

Chapter 2

Preliminaries

In this chapter we formally define PML, give necessary notations, and describe some basic properties that have been previously proved [OSVZ12, Zha05]. The notations used here may be different from those in published papers. We will be using or extending these properties and result in the other chapters.

- In Section 2.1 we define pattern-related notations.
- In Section 2.2 we briefly describe some known results.

2.1 Definitions

We define necessary notations that will be used throughout this dissertation.

2.1.1 Patterns

Let $\mathcal{A}$ be an alphabet of symbols. Let $\bar{x} = x_1 x_2 \cdots x_n$ be a sequence of symbols with $x_i \in \mathcal{A}$ for all $i \in [n] \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}$. The *pattern* of $\bar{x}$, denoted by $\psi(\bar{x})$, is a sequence of integers $\psi_1 \psi_2 \cdots \psi_n$ obtained from $\bar{x}$ by replacing each symbol with its order of appearance. More precisely, $\psi(\bar{x})$ is recursively defined as: $\psi_1 \stackrel{\text{def}}{=} 1$, and for all $i \geq 2$,

$$\psi_i \stackrel{\text{def}}{=} \begin{cases} \psi_j & \text{if } x_i = x_j \text{ for some } j \in [i-1], \\ \max\{\psi_1, \psi_2, \ldots, \psi_{i-1}\} + 1, & \text{otherwise.} \end{cases}$$

For example, $\psi(@) = 1, \psi(@\wedge) = 12, \psi(@ \wedge @) = 121$, and $\psi(alanpan) = 1213413$.

An integer sequence $\bar{\psi} = \psi_1\psi_2\cdots\psi_n$ is called a pattern if there exists at least once sequence $\bar{x}$ such that $\psi(\bar{x}) = \bar{\psi}$. Note that not all integer sequences are patterns. For example, 131 is not the pattern of any sequence. Given alphabet $\mathcal{A}$, a pattern of length n can be regarded as a subset of sequences in $\mathcal{A}^n \stackrel{\text{def}}{=} \{(a_1, a_2, \ldots, a_n) \mid a_i \in \mathcal{A}\}$. For example, for $\mathcal{A} = \{a, b, c\}$,

$$121 = \{aba, aca, bab, bcb, cac, cbc\}.$$

Let m be the largest number in $\bar{\psi}$, which is also the number of distinct symbols. For each $\psi \in [m]$, the *multiplicity* of ψ, denoted by μ_ψ, is the number of times ψ appear in $\bar{\psi}$. Let $\mathcal{M}(\bar{\psi})$ be the multiset of multiplicities. For any $\mu \in \mathcal{M}(\bar{\psi})$, the *prevalence* of μ, denoted by φ_μ, is the number of times μ appears in $\mathcal{M}(\bar{\psi})$.

For example, for the pattern $\bar{\psi} = 1213414$, $m = 4, \mu_1 = 3, \mu_2 = 1, \mu_3 = 1, \mu_4 = 2$, and $\mathcal{M} = \{3, 2, 1, 1\}^*$, where we use $\{\ldots\}^*$ to denote a *multiset*, and the prevalences are $\varphi_1 = 2, \varphi_2 = \varphi_3 = 1$.

For simplicity, if a number ψ appears i times consecutively, we abbreviate it as ψ^i. For example, we may write 11222111 in a shorter form $1^2 2^3 3^3$. A pattern of the form $1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m}$ such that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$ is *canonical*. For example, the canonical form of pattern 1213414 is $1^3 2^2 34$. Clearly any pattern has a unique canonical form.

2.1.2 Probabilities

Let P be a discrete distribution over alphabet $\mathcal{A}$. The probability of sequence $\bar{x} = x_1 x_2 \cdots x_n$ is

$$P(\bar{x}) = P(x_1) \cdot P(x_2 \mid x_1) \cdot \cdots \cdot P(x_n \mid x_1 x_2 \cdots x_{n-1}).$$

In this dissertation we consider only *i.i.d.* sequences. Then

$$P(\bar{x}) = P(x_1)P(x_2)\cdots P(x_n).$$

The probability of a pattern $\bar{\psi}$ is the total probability of all sequences in $\bar{\psi}$, *i.e.*,

$$P(\bar{\psi}) = \sum_{\bar{x} \in \bar{\psi}} P(\bar{x}).$$

For example, let $\mathcal{A} = \{a, b, c\}$ and $p_1 = P(a), p_2 = P(b), p_3 = P(c)$. Then

$$\begin{aligned} P(121) &= P(aba) + P(aca) + P(bab) + P(bcb) + P(cac) + P(cbc) \\ &= p_1^2 p_2 + p_1^2 p_3 + p_2^2 p_1 + p_2^2 p_3 + p_3^2 p_1 + p_3^2 p_2. \end{aligned}$$

Some observations:

- The pattern probability depends only on the multiplicities. For example, it is easy to see that $P(121) = P(112) = P(122)$, since they all have the same collection of multiplicities $\mathcal{M} = \{1, 1\}^*$. Thus as far as pattern probability is concerned we only need to consider canonical patterns.

- The pattern probability is a symmetric polynomial in the probabilities, thus it depends only on the multiset of probabilities, not how they are associated with symbols in $\mathcal{A}$. Let k be the support size of P. Then we may sort the probabilities and regard P as a vector in $\mathbb{R}^k$: $P = (p_1, p_2, \ldots, p_k)$ such that $p_1 \geq p_2 \geq \cdots \geq p_k \geq 0$ and $\sum_{i=1}^k p_i = 1$. When the context is clear, we use distribution and probability multiset interchangeably.

In general, given the distribution $P = (p_1, p_2, \ldots, p_k)$, the probability of the pattern $\bar{\psi} = 1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m}$ can be written as

$$P(\bar{\psi}) = \sum_{(i_1, i_2, \ldots, i_m) \in [k]^{\underline{m}}} p_{i_1}^{\mu_1} p_{i_2}^{\mu_2} \cdots p_{i_m}^{\mu_m}, \tag{2.1}$$

where $[k]^{\underline{m}} \stackrel{\text{def}}{=} \{(i_1, i_2, \ldots, i_m) \in [k]^m \mid i_t \neq i_{t'} \text{ for all } t \neq t'\}$, the set of vectors with distinct elements from $[k]$.

2.1.3 Pattern Maximum Likelihood

Let $\bar{x} = x_1 x_2 \cdots x_n$ be an *i.i.d.* sequence. Recall that *Sequence Maximum Likelihood* (SML) maximizes the probability of the sequence $\bar{x}$, *i.e.*,

$$P_{\text{SML}} \stackrel{\text{def}}{=} \arg\max_{P \in \mathcal{P}_d} P(\bar{x}),$$

where

$$\mathcal{P}_d \stackrel{\text{def}}{=} \left\{ (p_1, p_2, \ldots) \mid p_i \geq 0, \sum_{i=1}^{\infty} p_i = 1 \right\},$$

the set of all discrete distributions. For any $i \in [k]$, let

$$\mu_i(\bar{x}) \stackrel{\text{def}}{=} |\{j \mid x_j = a_i\}|,$$

the number of times the i-th symbol appears in $\bar{x}$. It can be shown that

$$P_{\text{SML}}(a_i) = \frac{\mu_i(\bar{x})}{n}.$$

Note that the support size of P_{SML} is always the number of distinct symbols in $\bar{x}$, and $P_{\text{SML}}(a) = 0$ for any $a \notin \{x_1, x_2, \ldots, x_n\}$.

Given pattern $\bar{\psi}$, the *Pattern Maximum Likelihood* (PML) of $\bar{\psi}$ is its largest possible probability over all discrete distributions:

$$\hat{P}(\bar{\psi}) \stackrel{\text{def}}{=} \max_{P \in \mathcal{P}_d} P(\bar{\psi}),$$

and the PML distribution of $\bar{\psi}$ achives the PML:

$$\hat{P}_{\bar{\psi}} \stackrel{\text{def}}{=} \arg\max_{P \in \mathcal{P}_d} P(\bar{\psi}).$$

As mentioned earlier, $P(\bar{\psi})$ does not depend on the association between probabilities and symbols. Denote the set of discrete distributions with sorted probabilities as

$$\mathcal{P}_d^{\text{sorted}} \stackrel{\text{def}}{=} \left\{ P = (p_1, p_2, \ldots) \mid p_1 \geq p_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} p_i = 1 \right\}$$

The following definitions are equivalent to the previous definitions:

$$\hat{P}(\bar{\psi}) \stackrel{\text{def}}{=} \max_{P \in \mathcal{P}_d^{\text{sorted}}} P(\bar{\psi}), \text{ and } \hat{P}_{\bar{\psi}} \stackrel{\text{def}}{=} \arg\max_{P \in \mathcal{P}_d^{\text{sorted}}} P(\bar{\psi}),$$

where $\hat{P}_{\bar{\psi}}$ differs from its previous definition by a permutation of the probabilities.

Note that in the domain of the maximization we are not restricting the support size of P to the number of observed distinct symbols m. We will show that there are indeed patterns maximized by distributions with supports size larger than m.

Examples:

(1) For pattern 1, any distribution P is a PML distribution since $P(1) \equiv 1$.

(2) For pattern 1^n, where $n > 1$, it is easy to see that the singleton distribution $P = (1)$ is the only one that can achieve the highest probability $P(1^n) = 1$. Thus $\hat{P}_{1^n} = (1)$.

(3) For pattern 112 and any distribution $P = (p_1, p_2, \ldots, p_k)$,

$$P(12) = \sum_{i\in[k]} \sum_{j\in[k]\setminus\{i\}} p_i^2 p_j = \sum_{i\in[k]} p_i^2(1-p_i) \le \sum_{i=1}^{k} p_i \cdot \frac{1}{4} = \frac{1}{4},$$

where the inequality follows from $p_i(1-p_i) \le \frac{1}{4}$. Clearly the maximum probability $\frac{1}{4}$ is achieved if and only if $p_i = \frac{1}{2}$ for all $i \in [k]$, namely $P = \left(\frac{1}{2}, \frac{1}{2}\right)$.

(4) Consider the pattern 12. For any discrete distribution $P = (p_1, p_2, \ldots, p_k)$,

$$P(12) = \sum_{i\in[k]} \sum_{j\in[k]\setminus\{i\}} p_i p_j = \sum_{i\in[k]} p_i(1-p_i) = 1 - \sum_{i\in[k]} p_i^2.$$

Since $\sum_{i\in[k]} p_i^2$ is strictly greater than 0, $P(12)$ can never achieve 1. However, if $p_1 = p_2 = \cdots = p_k$,

$$\sum_{i\in[k]} p_i^2 = k \cdot \frac{1}{k^2} = \frac{1}{k},$$

which goes to 0 as k goes to infinity, and hence $P(12)$ goes to, but never equal to, 1. This means that $\hat{P}_{12}$ is not well-defined over discrete distributions.

To ensure the existence of PML distributions, we modify the domain of distributions as follows. Instead of discrete distributions, we consider *mixture distributions* that have both discrete part and a continuous part. Note that the pattern probability depends on only the total probability in the continuous part; the density function does not matter. Thus we can represent a mixture distribution P as a vector of real values $P = (p_1, p_2, \ldots, p_k)$, where $p_1 \ge p_2 \ge \cdots \ge p_k > 0$ and $\sum_{i=1}^{k} p_i \le 1$; the continuous probability q is implicit:

$$q \stackrel{\text{def}}{=} 1 - \sum_{i=1}^{k} p_i.$$

The probability of a pattern underlying a mixture distribution can be calculated similarly. For example, let $P = (p_1, p_2, p_3)$ be a mixture distribution with continuous probability $q = 1 - (p_1 + p_2 + p_3)$. Then

$$P(112) = (p_1^2 p_2 + p_1^2 p_3 + p_2^2 p_1 + p_2^2 p_3 + p_3^2 p_1 + p_3^2 p_2) + (p_1^2 q + p_2^2 q + p_3^2 q).$$

Recall that φ_1 is the number of multiplicities equal to 1, *i.e.*

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m - \varphi_1 > \mu_{m-\varphi_1+1} = \cdots = \mu_m = 1.$$

Then, for any mixture distribution $P = (p_1, p_2, \ldots, p_k)$,

$$P(\bar{\psi}) = \sum_{\ell=0}^{\varphi_1} \binom{\varphi_1}{\ell} q^\ell \sum_{(i_1, i_2, \ldots, i_{m-\ell}) \in [k]^{\underline{m-\ell}}} p_{i_1}^{\mu_1} p_{i_2}^{\mu_2} \cdots p_{i_{m-\ell}}^{\mu_{m-\ell}} \tag{2.2}$$

With the modified definition, it is easy to see that $\hat{P}_{12} = (\,)$, *i.e.*, the PML distribution has no discrete part and the continuous part has probability $q = 1$. Let

$$\mathcal{P}_{\text{mix}}^{\text{sorted}} = \left\{ (p_1, p_2, \ldots) \mid p_1 \geq p_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} p_i \leq 1 \right\},$$

the set of all mixture distributions with sorted discrete probabilities. Then the following modified definitions are well-defined:

$$\hat{P}(\bar{\psi}) \stackrel{\text{def}}{=} \max_{P \in \mathcal{P}_{\text{mix}}^{\text{sorted}}} P(\bar{\psi}), \text{ and } \hat{P}_{\bar{\psi}} \stackrel{\text{def}}{=} \arg\max_{P \in \mathcal{P}_{\text{mix}}^{\text{sorted}}} P(\bar{\psi}).$$

Occasionally, we need to consider only discrete distributions. To allow the existence of maxima, we need to bound the support size. Given pattern $\bar{\psi}$, let $K \geq m$ be an upper bound on the support size. Then the following *bounded* PML is also well-defined:

$$\hat{P}^{(K)}(\bar{\psi}) \stackrel{\text{def}}{=} \max_{P \in \mathcal{P}_d^{\text{sorted}}: k \leq K} P(\bar{\psi}), \text{ and } \hat{P}_{\bar{\psi}}^{(K)} \stackrel{\text{def}}{=} \arg\max_{P \in \mathcal{P}_d^{\text{sorted}}: k \leq K} P(\bar{\psi}),$$

where k is the support size of P. It can be shown, by continuity, that

$$\lim_{K \to \infty} \hat{P}^{(K)}(\bar{\psi}) = \hat{P}(\bar{\psi}).$$

Table 2.1 summaries some of the notations we will be using frequently.

Table 2.1: Notations

symbol	description
ψ	pattern
n	pattern length
m	number of distinct symbols
μ_t	multiplicity
φ_μ	prevalence
μ_S	summation of multiplicities with indices in S
$\bar{\psi} \otimes \bar{\psi}'$	concatenation
P, $\hat{P}$	distribution (probability multiset), PML distribution
$\hat{P}^{(K)}$	PML distribution with bounded support size
p_i, $\hat{p}_i$	discrete probability, PML probability
q, $\hat{q}$	continuous probability, PML continuous probability
k, $\hat{k}$	support size, PML support size
$\mathcal{P}_d^{\text{sorted}}$	set of discrete distributions with sorted probabilities
$\mathcal{P}_{\text{mix}}^{\text{sorted}}$	set of mixture distributions with sorted discrete probabilities
$P_i(\bar{\psi})$	defined as $\sum_{\bar{x}\in\bar{\psi}:\, a_i\notin\bar{x}} P(\bar{x})$

2.1.4 More notations

We introduce additional notations that we will use in most chapters. Let $\bar{\psi} = 1^{\mu_1}2^{\mu_2}\cdots m^{\mu_m}$ be a canonical pattern. Given a subset $S \subseteq [m]$, the *sub-pattern* of $\bar{\psi}$ restricted to S, denoted by $\bar{\psi}_S$, is the canonical pattern obtained by removing integers not in S from $\bar{\psi}$, and then re-label the remaining ones. In other words, $\bar{\psi}_S$ has multiplicities

$$\mathcal{M}(\bar{\psi}_S) = \{\mu_t \mid t \in S\}.$$

For example, given pattern $1^3 2^2 3^2 45$ and $S = \{1, 3, 4\}$, we have $\mathcal{M}(\bar{\psi}_S) = \{3, 2, 1\}^*$, hence $\bar{\psi}_S = 1^3 2^2 3$.

For simplicity, if S contains all indices in $[m]$ except for a few, we use the missing indices in the subscript. For example,

$$\bar{\psi}_t \stackrel{\text{def}}{=} \bar{\psi}_{[m]\setminus\{t\}}, \quad \bar{\psi}_{t_1,t_2} \stackrel{\text{def}}{=} \bar{\psi}_{[m]\setminus\{t_1,t_2\}},$$

and so on.

Given two patterns $\bar{\psi} = 1^{\mu_i}2^{\mu_i}\cdots m^{\mu_i}$ and $\bar{\psi}' = 1^{\mu'_1}2^{\mu'_2}\cdots m'^{\mu'_{m'}}$, their *concatenation* is

$$\bar{\psi} \otimes \bar{\psi}' \stackrel{\text{def}}{=} 1^{\mu_i}2^{\mu_i}\cdots m^{\mu_i}(m+1)^{\mu'_1}(m+2)^{\mu'_2}\cdots(m+m')^{\mu'_{m'}}.$$

Given a list of ordered values (real numbers or integers) $V = (V_1, V_2, \ldots)$, and a set of indices $S \subseteq \mathbb{N} \stackrel{\text{def}}{=} \{1, 2, \ldots\}$, let

$$V_S \stackrel{\text{def}}{=} \sum_{i \in S} V_i.$$

For example, for multiplicities $(\mu_1, \mu_2, \ldots, \mu_m)$, we use μ_S for $\sum_{i \in S} \mu_i$. For a distribution $(p_1, p_2, \ldots, p_k)$, we use p_S for $\sum_{i \in S} p_i$.

Given pattern $\bar{\psi}$, a distribution $P = (p_1, p_2, \ldots, p_k)$ over an alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$, and a set of indices $I \subseteq [k]$, let $P_I(\bar{\psi})$ be the probability that a sequence has pattern $\bar{\psi}$ and it has only symbols from $\mathcal{A}_I \stackrel{\text{def}}{=} \{a_i \mid i \in I\}$. More precisely,

$$P_I(\bar{\psi}) \stackrel{\text{def}}{=} \sum_{\bar{x} \in \mathcal{A}_I : \psi(\bar{x}) = \bar{\psi}} P(\bar{x}).$$

For simplicity, when I contains all indices in $[k]$ except a few, we use the missing indices in the subscript instead. For example,

$$P_i(\bar{\psi}) \stackrel{\text{def}}{=} P_{[k] \setminus \{i\}}(\bar{\psi}), \quad P_{i,j}(\bar{\psi}) \stackrel{\text{def}}{=} P_{[k] \setminus \{i,j\}}(\bar{\psi})$$

and so on.

Given a pattern $\bar{\psi} = 1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m}$ and a distribution $P \in \mathcal{P}_d^{\text{sorted}}$, the pattern probability can be expanded in two ways:

(1) For a given $t \in [m]$, by considering what the first symbol is in $\bar{\psi}$, we can write the pattern probability as

$$P(\bar{\psi}) = \sum_{i=1}^{k} p_i^{\mu_t} P_i(\bar{\psi}_t) + I_{\mu_t = 1} \cdot q P(\bar{\psi}_m), \tag{2.3}$$

where $I_{\mu_t = 1}$ is the $\{0, 1\}$ indicator function.

(2) For a given $i \in [k]$, by considering the number of times the i-th symbol appears, we have

$$P(\bar{\psi}) = P_i(\bar{\psi}) + \sum_{t=1}^{m} p_i^{\mu_t} P_i(\bar{\psi}_t). \tag{2.4}$$

We will be using both Expansions (2.3) and (2.4) frequently. More generally, for any $S \subseteq [m]$,

$$P(\bar{\psi}) = \sum_{I \subseteq [k]} P_I(\bar{\psi}_S) P_{\bar{I}}(\bar{\psi}_{\bar{S}}), \tag{2.5}$$

where $\bar{I} = [k] \setminus I$ and $\bar{S} = [m] \setminus S$. On the other hand, given any $I \subseteq [k]$,

$$P(\bar{\psi}) = \sum_{S \subseteq [m]} P_I(\bar{\psi}_S) P_{\bar{I}}(\bar{\psi}_{\bar{S}}). \tag{2.6}$$

Sometimes we also need to expand $P(\bar{\psi})$ by the continuous probability q. By considering which *singletons*, symbols that appear only once, are from the continuous part, we have

$$P(\bar{\psi}) = \sum_{\ell=0}^{\varphi_1} \binom{\varphi_1}{\ell} q^\ell P_q(\bar{\psi}_{[1..m-\ell]}), \tag{2.7}$$

where $P_q(\cdot)$ is the probability of a pattern with no symbols from the continuous part.

2.2 Previous results

We briefly describe some of the known results in [OSVZ12, Zha05] that we will be using or extending in the remaining chapters. To distinguish existing results from the new results that we will show in this dissertation, we use **Fact** for the existing results.

To get started, we first answer the following two questions:

(1) Does the PML distribution always exist?

(2) Does the PML distribution converge to the underlying distribution?

The answers to both questions are positive.

2.2.1 Existence

Note that the PML distribution is defined as a maximization problem over the domain $\mathcal{P}_{\text{mix}}^{\text{sorted}}$. The existence of the maximum is guarantee by the well-known

Extreme Value Theorem, which states that the maximum/minimum of a continuous function can be achieved in a compact space. The compactness of $\mathcal{P}_{\text{mix}}^{\text{sorted}}$ can be proved by showing that it is both complete (every Cauchy sequence converges) and totally bounded (it can be covered by finitely many open balls of any fixed size). For any pattern $\bar{\psi}$, the continuity of $P(\bar{\psi})$ as a function of $P \in \mathcal{P}_{\text{mix}}^{\text{sorted}}$ can be proved inductively on m, the number of distinct symbols in $\bar{\psi}$. Thus the existence of $\hat{P}_{\bar{\psi}}$ follows from the Extreme Value Theorem:

Fact 2.1. *For all patterns $\bar{\psi}$, the maximum of $P(\bar{\psi})$ over mixed distributions can be achieved.*

It remains open whether the PML distribution of any pattern is always unique, although no example with two different PML distributions has been found except for the trivial pattern 1 whose PML distribution is any distribution. Unless otherwise specified, all the properties mentioned in this dissertation apply to any PML distribution, if not unique.

2.2.2 Consistency

Let $x_1^n = x_1 x_2 \cdots x_n$ be an *i.i.d.* sequence of length n drawn from an unknown discrete distribution $P \in \mathcal{P}_d$. An *estimator* f_n is a function that estimates P upon observing x_1^n. A sequence of estimators $\{f_n\}$ is *consistent* if f_n converges to P in probability. In other words, $\{f_n\}_{n=1}^{\infty}$ is consistent if for all $\epsilon > 0$ and all $P \in \mathcal{P}_d$,

$$\lim_{n\to\infty} \Pr(\|f_n - P\| > \epsilon) = 0,$$

where the norm $\|\cdot\|$ represents the metric of interest. The sequence $\{f_n\}_{n=1}^{\infty}$ is *uniformly consistent*, if for all $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_d} \Pr(\|f_n - P\| > \epsilon) = 0.$$

The definition of consistency can be refined to incorporate the notation of the rate of convergence. The sequence $\{f_n\}_{n=1}^{\infty}$ is *uniformly k_n-consistent* if there exists $M > 0$ such that

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_d} \Pr(k_n \cdot \|f_n - P\| > M) = 0.$$

For any $P, Q \in \mathcal{P}_d$, define the L_∞ distance as $\|P-Q\|_\infty = \max_{i=1}^{\infty}\{p_i - q_i\}$. It can be shown that the PML estimator $\hat{P}_n \stackrel{\text{def}}{=} \hat{P}_{\psi(x_1^n)}$ is consistent under L_∞ distance:

Fact 2.2. *The PML estimator $\{\hat{P}\}$ is uniformly $\frac{n^{1/4}}{\ln n}$-consistent for $P \in \mathcal{P}_d^{\text{sorted}}$ with respect to L_∞.*

2.2.3 Majorization

Given two distributions $P = (p_1, p_2, \ldots) \in \mathcal{P}_d^{\text{sorted}}$ and $Q = (q_1, q_2, \ldots) \in \mathcal{P}_d^{\text{sorted}}$, P *majorizes* Q, written as $P \succeq Q$ or $Q \preceq P$, if for all $i \geq 1$,

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

Intuitively, Q is "smoother" than P. Note that any distribution in $\mathcal{P}_d^{\text{sorted}}$ majorizes the uniform distribution of the same support size. Roughly speaking, uniform distributions are the "smoothest". It can be shown for any pattern the PML distribution is always smoother than the SML distribution:

Fact 2.3. *For all patterns $\bar{\psi} = 1^{\mu_1}2^{\mu_2}\cdots m^{\mu_m}$,*

$$P_{\text{SML}} \succeq \hat{P}_{\bar{\psi}},$$

where $P_{\text{SML}} \stackrel{\text{def}}{=} \left(\frac{\mu_1}{n}, \frac{\mu_2}{n}, \ldots, \frac{\mu_m}{n}\right)$.

We can define majorization between patterns in a similar way. Let $\bar{\psi} = 1^{\mu_1}2^{\mu_2}\cdots m^{\mu_m}$ and $\bar{\psi}' = 1^{\mu'_1}2^{\mu'_2}\cdots m^{\mu'_m}$ be canonical patterns with equal length $n = n'$ and the same number of distinct symbols $m = m'$. Then $\bar{\psi}$ *majorizes* $\bar{\psi}'$, written $\bar{\psi} \succeq \bar{\psi}'$ or $\bar{\psi}' \preceq \bar{\psi}$, if for all $i \in [m]$,

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

For any discrete distribution $P \in \mathcal{P}_d$, it follows directly from Muirhead's Inequality [Mui02] that $P(\bar{\psi}) \geq P(\bar{\psi}')$. For a mixture distribution $P \in \mathcal{P}_{\text{mix}}^{\text{sorted}}$, the same inequality follows from the continuity of $P(\bar{\psi})$ as a function of P and that $\mathcal{P}_{\text{mix}}^{\text{sorted}}$

is complete. It immediately follows that $\hat{P}(\bar{\psi}) \geq \hat{P}_{\bar{\psi}}(\bar{\psi}') \geq \hat{P}(\bar{\psi}')$.

Fact 2.4. *For any pattern $\bar{\psi}$ and $\bar{\psi}'$ such that $\bar{\psi} \succeq \bar{\psi}'$,*

$$\hat{P}_{\bar{\psi}}(\bar{\psi}) \geq \hat{P}_{\bar{\psi}'}(\bar{\psi}).$$

Given n and m, let $\mu = \lceil n/m \rceil, \varphi = \mu m - n$. The patterns

$$\bar{\psi}^{\uparrow} \stackrel{\text{def}}{=} 1^{n-m+1}23\cdots m \text{ and}$$
$$\bar{\psi}^{\downarrow} \stackrel{\text{def}}{=} 1^{\mu}2^{\mu}\cdots(m-\varphi)^{\mu}(m-\varphi+1)^{\mu-1}\cdots m^{\mu-1}$$

are called *skewed* and *1-uniform* patterns, respectively. it is easy to see that $\bar{\psi}^{\uparrow}$ majorizes, while $\bar{\psi}^{\downarrow}$ is majorized by, all other patterns with the same length and number of distinct symbols. Thus, the PML of any pattern $\bar{\psi}$ can be bounded using skewed and 1-uniform patterns:

$$\hat{P}(\bar{\psi}^{\uparrow}) \geq \hat{P}_{\bar{\psi}}(\bar{\psi}) \geq \hat{P}(\bar{\psi}^{\downarrow}).$$

We'll further study skewed patterns and 1-uniform patterns Chapter 4.

2.2.4 Continuous Probability

As we have seen before, given pattern $\bar{\psi}$, its PML distribution $\hat{P}_{\bar{\psi}}$ may have positive continuous probability $\hat{q} > 0$. Recall that φ_1 is the number of symbols in $\bar{\psi}$ that appear only once. It can be shown that $\hat{q}$ can be bounded using φ_1:

Fact 2.5. *For all patterns $\bar{\psi} \neq 1$,*

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

In a *singleton-free pattern* $\varphi_1 = 0$, In a *unique-singleton pattern* $\varphi_1 = 1$, *i.e.*, no symbols appear once. A direct consequence of Fact 2.5 is that the PML distribution of any singleton-free pattern is discrete. In fact, we can show that this also holds for unique-singleton patterns:

Fact 2.6. *The PML distribution is dicrete,* i.e.*, $q = 0$, for any pattern $\bar{\psi} \neq 1$ such that $\varphi_1 \leq 1$.*

2.2.5 Support Size

Given pattern $\bar{\psi}$, it can be shown that the discrete support size $\hat{k}$ of the PML distribution must be finite, by bounding the the number of distinct PML probabilities:

Fact 2.7. *For all patterns $\bar{\psi} \neq 1$,*

$$|\{\hat{p}_1, \hat{p}_2, \ldots\}| \leq \min\{n-1, 2^m\}.$$

Furthermore, if the smallest multiplicity μ_m is greater than 1, $\hat{k}$ can be bounded from both below and above:

Fact 2.8. *For all patterns $\bar{\psi} \neq 1$ such that $\hat{q} = 0$,*

$$m - 1 + \frac{\sum_{i \in [2\,..\,m]} 2^{-\mu_i}}{2^{\mu_1} - 2} \leq \hat{k} \leq m + \frac{m-1}{2^{\mu_m} - 2}.$$

It is easy to see that by Fact 2.8, if $\mu_m > \log_2(m+1)$, then $\hat{k} = m$, *i.e.*, there is no unseen symbol in the PML estimate.

2.2.6 Patterns with Known PML

We have seen that for some very simple patterns the PML distribution can be found analytically. We describe a list of patterns whose PML distributions were previous found. These results will be extended in Chapter 4.

The patterns 1^n and $12\cdots n$ are *trivial*. It is easy to see that

Fact 2.9. *$\hat{P}_{1^n} = (1)$, $\hat{P}_1$ is any distribution, and $\hat{P}_{12\cdots n} = (\,)$ for all $n \geq 2$.*

In a *uniform* pattern all the multiplicities are the same, *i.e.*, $\mu_1 = \cdots = \mu_m$. The PML distribution of a non-trivial uniform pattern is always uniform. Its support size $\hat{k}$ can be found as follows. Let $P = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$. Then

$$P(\bar{\psi}) = f(k) \stackrel{\text{def}}{=} k^{\underline{m}} \left(\frac{1}{k}\right)^n,$$

where $k^{\underline{m}} \stackrel{\text{def}}{=} k(k-1)\cdots(k-m+1)$. Note that

$$\frac{f(k+1)}{f(k)} = \frac{k+1}{k-m+1} \cdot \left(\frac{k}{k+1}\right)^n,$$

which can be shown to have a unique maxima. Then $\hat{k}$ can be found as the smallest k such that

$$\frac{k+1}{k-m+1} \cdot \left(\frac{k}{k+1}\right)^n \leq 1.$$

Fact 2.10. *For any $\mu \geq 2$, $\hat{P}_{1^\mu 2^\mu \cdots m^\mu}$ is uniform with support size*

$$\hat{k} = \arg\min_{k \geq m} \frac{k+1}{k-m+1} \cdot \left(\frac{k}{k+1}\right)^n \leq 1.$$

In a *1-uniform* pattern the multiplicities differ by at most 1, *i.e.*$\mu_1 - \mu_m \leq 1$. It was shown that the PML of all non-trivial 1-uniform patterns can be achieved at a uniform distribution. The support size can be found in the same way as for uniform patterns.

Fact 2.11. *The PML of any 1-uniform pattern can be achieved at a uniform distribution with support size*

$$\hat{k} = \arg\min_{k \geq m} \frac{k+1}{k-m+1} \cdot \left(\frac{k}{k+1}\right)^n \leq 1.$$

A *binary* pattern $1^{\mu_1}2^{\mu_2}$ has two distinct symbols, *i.e.*, $m = 2$. It was shown that the PML distribution of any non-trivial binary pattern is discrete with support size 2. Let $\hat{P}_{1^{\mu_1}2^{\mu_2}} = (p, 1-p)$. Then

$$\hat{P}(1^{\mu_1}2^{\mu_2}) = \max_{\frac{1}{2} \leq p < 1} p^{\mu_1}(1-p)^{\mu_2} + (1-p)^{\mu_1}p^{\mu_2}.$$

It is easy to see that the optimal p can be obtained by solving the equation

$$\frac{d}{dp}\left[p^{\mu_1}(1-p)^{\mu_2} + (1-p)^{\mu_1}p^{\mu_2}\right] = 0,$$

which can be rewritten as $\left(\frac{p}{1-p}\right)^{\mu_1-\mu_2} + \frac{\mu_2 - np}{\mu_1 - np} = 0$.

Fact 2.12. *The PML distribution of any non-trivial binary pattern $1^{\mu_1}2^{\mu_2}$ is discrete with support size 2, and the probabilities $\hat{p}_i$, $i = 1, 2$, can be found by solving the equation*

$$\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)^{\mu_1-\mu_2} + \frac{\mu_2 - n\hat{p}_i}{\mu_1 - n\hat{p}_i} = 0.$$

Furthermore, if $(\mu_1 - \mu_2)^2 \leq n$, then $\hat{P}_{1^{\mu_1}2^{\mu_2}}$ is uniform, i.e., $\hat{p}_1 = \hat{p}_2 = \frac{1}{2}$.

In a *skewed* pattern $\bar{\psi} = 1^r 23\cdots(u+1)$ one symbol repeats r times and the other u symbols are unique. For $r = 1$, $\bar{\psi}$ degenerates to a trivial pattern. For $r = 2$, $\bar{\psi}$ becomes 1-uniform. For $u = 1$, $\bar{\psi}$ is binary.

The other cases are *truly skewed*, *i.e.*, $r \geq 3$ and $u \geq 2$. The conjecture for a truly skewed pattern is that

Conjecture 1. *For any $r \geq 3$ and $u \geq 2$,*

$$\hat{P}_{1^r 23\cdots(u+1)} = \left(\frac{r}{r+u}\right).$$

That is, the PML distribution of a truly skewed pattern has only a single discrete probability $\hat{p}_1 = \frac{r}{r+u}$; all the other probability $\hat{q} = \frac{u}{r+u}$ goes to the continuous part. This conjecture was shown to be true asymptotically:

Fact 2.13. *For u sufficiently large and $r \geq 2\sqrt{u}$,*

$$\hat{P}_{1^r 23\cdots(u+1)} = \left(\frac{r}{r+u}\right).$$

We will show in Section 4.2 of Chapter 4 that the conjecture holds for all truly skewed patterns.

2.2.7 Approximation Algorithm

For general pattern finding the exact PML distribution may be difficulty. An EM algorithm was proposed [OSS+04, Zha05]. We will describe the algorithm in details in Chapter 6, and show that it is equivalent to a Generalized Gradient Ascend method. We will use the algorithm to evaluate the performance of PML on various distributions, and apply it to the authorship authentication of the Taylor poem. In Chapter 7 We will extend the algorithm to set-patterns.

Chapter 3

Properties of PML

Chapter 2 described some basic properties of PML. We further extend the results on the support size and continuous probability.

- In Section 3.1 we prove a few useful (in)equalities concerning the partial derivatives of the pattern probability with respect to the probabilities.
- In Section 3.2 we extend previous bounds on the discrete PML support size. In addition, we show upper bounds on the number of identical probabilities.
- In Section 3.3 we show a larger class of patterns whose PML distribution has no continuous part.

3.1 Partial Derivatives

We first give some (in)equalities that we will be frequently using. Recall that, as given in Expansion (2.2), the pattern probability can be regarded as a multi-variate function of the discrete probabilities $p_1, p_2, \ldots, p_k$ and the continuous probability q. We show the following relations between the partial derivatives.

Lemma 3.1. *Let $P = (p_1, p_2, \ldots, p_k) = \hat{P}_{\bar{\psi}} \in \mathcal{P}_{\text{mix}}^{\text{sorted}}$ be the PML distribution of a non-trivial pattern $\bar{\psi}$. Then, for any $i \in [k]$,*

$$\frac{\partial P(\bar{\psi})}{\partial p_i} = nP(\bar{\psi}) \geq \frac{\partial P(\bar{\psi})}{\partial q} = \varphi_1 P(\bar{\psi}_m),$$

where the second equality holds if $q > 0$.

Proof. Maximizing $P(\bar{\psi})$ is equivalent to maximizing $\ln P(\bar{\psi})$. Using Lagrangian multiplier, we remove the constraint $\sum_{i=1}^{k} p_k + q = 1$ by maximizing

$$f(P, \lambda) \stackrel{\text{def}}{=} \ln P(\bar{\psi}) + \lambda \left(1 - q - \sum_{i=1}^{k} p_i\right).$$

It follows that $P = \hat{P}_{\bar{\psi}}$ satisfies, for all $i \in [k]$,

$$\begin{aligned}
\frac{\partial f}{\partial p_i}(P, \lambda) &= \frac{1}{P(\bar{\psi})} \frac{\partial P(\bar{\psi})}{\partial p_i} - \lambda = 0, \\
\frac{\partial f}{\partial q}(P, \lambda) &= \frac{1}{P(\bar{\psi})} \frac{\partial P(\bar{\psi})}{\partial q} - \lambda \leq 0,
\end{aligned}$$

where the last equality holds if the optimal $q = \hat{q}$ is positive. Thus

$$\frac{\partial P(\bar{\psi})}{\partial p_1} = \frac{\partial P(\bar{\psi})}{\partial p_2} = \cdots = \frac{\partial P(\bar{\psi})}{\partial p_k} = \lambda P(\bar{\psi}) \geq \frac{\partial P(\bar{\psi})}{\partial q}.$$

By Expansion (2.7),

$$\begin{aligned}
\frac{\partial P(\bar{\psi})}{\partial q} &= \sum_{\ell=1}^{\varphi_1} \binom{\varphi_1}{\ell} \ell q^{\ell-1} P_q(\bar{\psi}_{[1\,..\,m-\ell]}) \\
&= \varphi_1 \sum_{\ell=1}^{\varphi_1} \binom{\varphi_1 - 1}{\ell - 1} q^{\ell-1} P_q(\bar{\psi}_{[1\,..\,m-\ell]}) \\
&= \varphi_1 P(\bar{\psi}_m).
\end{aligned}$$

Thus it remains to show that the optimal λ is n. To see this, note that from Expansion (2.4) that

$$p_i \frac{\partial P(\bar{\psi})}{\partial p_i} = \sum_{t=1}^{m} \mu_t p_i^{\mu_t} P_i(\bar{\psi}_t).$$

There are two cases.

- If $q = 0$, then

$$\lambda = \sum_{i=1}^{k} p_i \lambda = \frac{1}{P(\bar{\psi})} \sum_{i=1}^{k} p_i \frac{\partial P(\bar{\psi})}{\partial p_i} = \frac{1}{P(\bar{\psi})} \sum_{t=1}^{m} \mu_t \sum_{i=1}^{k} p_i^{\mu_t} P_i(\bar{\psi_t}).$$

  On the other hand, since $q = 0$, it follows from Expansion (2.3) that, for any $t \in [m]$, $\sum_{i=1}^{k} p_i^{\mu_t} P_i(\bar{\psi_t}) = P(\bar{\psi})$. Thus

$$\lambda = \frac{1}{P(\bar{\psi})} \cdot \sum_{t=1}^{m} \mu_t P(\bar{\psi}) = n.$$

- If $q > 0$, it follows from Expansion (2.7) that

$$q \frac{\partial P(\bar{\psi})}{\partial q} = \sum_{S\colon \forall t \in S,\, \mu_t = 1} |S| q^{|S|} P_q(\bar{\psi}_{\bar{S}}),$$

  which can be rewritten as

$$q \frac{\partial P(\bar{\psi})}{\partial q} = \sum_{t\colon \mu_t = 1} \sum_{S \ni t\colon \forall t' \in S,\, \mu_{t'} = 1} q^{|S|} P_q(\bar{\psi}_{\bar{S}}) = \sum_{t\colon \mu_t = 1} q P(\bar{\psi_t}).$$

  Then

$$\begin{aligned}
\lambda &= \left(\sum_{i=1}^{k} p_i + q \right) \lambda \\
&= \frac{1}{P(\bar{\psi})} \left[\sum_{i=1}^{k} p_i \frac{\partial P(\bar{\psi})}{\partial p_i} + q \frac{\partial P(\bar{\psi})}{\partial q} \right] \\
&= \frac{1}{P(\bar{\psi})} \sum_{t=1}^{m} \mu_t \left[\sum_{i=1}^{k} p_i^{\mu_t} P_i(\bar{\psi_t}) + \sum_{t\colon \mu_t = 1} q P(\bar{\psi_t}) \right].
\end{aligned}$$

  It follows from Expansion (2.3) that

$$\lambda = \frac{1}{P(\bar{\psi})} \sum_{t=1}^{m} \mu_t P(\bar{\psi}) = n.$$

□

3.2 PML Support Size

In Chapter 2 we introduced previous results on the support size of PML distributions. In Fact 2.8, it was shown that for PML distributions with no continuous part,

$$m - 1 + \frac{\sum_{i \in [2\,..\,m]} 2^{-\mu_i}}{2^{\mu_1} - 2} \leq \hat{k} \leq m + \frac{m-1}{2^{\mu_m} - 2}.$$

Note that the upper bound is useful only if $\mu_m > 1$. In this section we find bounds that can be applied to more patterns.

We will first show the following two lower bounds on $\hat{k}$.

Theorem 3.2. *For any pattern $\bar{\psi}$, if $\hat{q} = 0$, then*

$$\hat{k} \geq \frac{\binom{\varphi_1}{2}}{m - \varphi_1} + m - 1,$$

where the equality holds only if $\hat{P}_{\bar{\psi}}$ is uniform.

Theorem 3.3. *For any pattern $\bar{\psi}$, if $\hat{q} = 0$, then*

$$\hat{k} \geq m - 1 + \frac{\sum_{\{t,t'\}} 2^{-(\mu_t + \mu_{t'} - 1)}}{m - \sum_t 2^{-(\mu_t - 1)}}.$$

For some patterns the lower bound in Theorem 3.3 is better than that in Theorem 3.2. Furthermore, we can derive a simpler lower bound from Theorem 3.3 as follows, which is although slighter weaker than that in Thmosition 3.2. For any $t \in [m]$,

$$\frac{\sum_{\{t,t'\}} 2^{-(\mu_t + \mu_{t'} - 1)}}{m - \sum_t 2^{-(\mu_t - 1)}} = \frac{2^{-(\mu_t - 1)} \cdot \sum_{t' \neq t} 2^{-\mu_{t'}} + \sum_{\{t',t''\}: t',t'' \neq t} 2^{-(\mu_{t'} + \mu_{t''} - 1)}}{1 - 2^{-(\mu_t - 1)} + \sum_{t' \neq t} (1 - 2^{-(\mu_{t'} - 1)})},$$

which decreases in μ_t. Let $\mu_t \to \infty$ for all $\mu_t \geq 2$. Then

$$\frac{\sum_{t,t'} 2^{-(\mu_t + \mu_{t'} - 1)}}{m - \sum_t 2^{-(\mu_t - 1)}} \geq \frac{\frac{1}{2}\binom{\varphi_1}{2}}{m - \varphi_1}.$$

We will also show the following upper bounds.

Theorem 3.4. *For any pattern such that $c = \frac{2(m-1)\varphi_2}{(n-1)\varphi_1} > 1$,*

$$\hat{k} \leq \frac{c(n-1)}{c-1}.$$

In a *unique-singleton pattern* $\varphi_1 = 1$. We will show that the PML support size in linear in m for all unique-singleton patterns.

Theorem 3.5. *For any pattern such that $\varphi_1 = 1$,*

$$\hat{k} < m + \frac{m-1}{\mu^*_{m-1}},$$

where

$$\mu^*_{m-1} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } \mu_{m-1} = 2, \\ \mu_{m-1} & \text{otherwise.} \end{cases}$$

Let

$$\mu_* \stackrel{\text{def}}{=} \mu_{m-\varphi_1} = \min\{\mu_t : \mu_t \geq 2\},$$

the smallest multiplicity that is greater than one. Furthermore, let

$$\bar{\psi}_* \stackrel{\text{def}}{=} \bar{\psi}_{m-\varphi_1},$$

the pattern obtained from $\bar{\psi}$ with the symbol corresponding to μ_* times removed. In a *quasi-skewed pattern* $prev_2 = 0$, *i.e.*, $\mu_* \geq 3$. We will prove the following upper bounds on the PML support size for quasi-skewed patterns.

Theorem 3.6. *For any pattern $\bar{\psi}$ such that $\varphi_2 = 0$,*

$$\hat{k} \leq \sqrt[\mu_*-2]{\frac{m-\varphi_1}{\binom{\varphi_1}{2}} \cdot \frac{\hat{P}(\bar{\psi}_*)}{\hat{P}(\bar{\psi})}}.$$

Note that In Theorem 3.6 although the exact values of $\hat{P}(\bar{\psi}_*)$ and $\hat{P}(\bar{\psi})$ may be unknown, we can get a finite upper bound on $\hat{k}$ by replacing $\hat{P}(\bar{\psi}_*)$ and $\hat{P}(\bar{\psi})$ with an upper bound and a lower bound respectively. We will show that

Corollary 3.7. *For any pattern $\bar{\psi}$ such that $\mu_* \geq 3$,*

$$\hat{k} \leq \sqrt[\mu_*-2]{\frac{m-\varphi_1}{\binom{\varphi_1}{2}} e^{2n}}.$$

In a *singleton-free pattern* $\varphi_1 = 0$. Given $P = (p_1, p_2, \ldots, p_k) \in \mathcal{P}^{\text{sorted}}_{\text{mix}}$, let $\nu(p_i)$ be the number of probabilities in P that are equal to p_i. Using a different approach we will also show the following upper bounds for singleton-free patterns and quasi-skewed patterns:

Theorem 3.8. *The PML distribution of any pattern $\bar{\psi} \neq 1$ satisifies*

- *If $\varphi_1 = 0$, then,for all $i \in [\hat{k}]$,*

$$\nu(\hat{p}_i) \leq \max_{S \subseteq [m]: n_S - m_S \geq 1} \frac{(n_S - 1)(m_S - 1)}{n_S - m_S} + 1,$$

  which implies that $\nu(\hat{p}_i) \leq 2m$.

- *If $\varphi_2 = 0$, then, for all $i \in [\hat{k}]$,*

$$\nu(\hat{p}_i) \leq \max_{S \subseteq [m]: n_S - m_S \geq 2} \frac{(n_S - 1)(m_S - 1)}{n_S - m_S - 1},$$

  which implies that $\nu(\hat{p}_i) \leq m^2$ and

$$\hat{k} \leq m^2 \cdot \min\{n - 1, 2^m\}.$$

Theorem 3.2: First Lower Bound

Proof of Theorem 3.2. For simplicity, let $P = (p_1, p_2, \ldots, p_k) = \hat{P}_{\bar{\psi}}$. For any $i \in [k]$, let $P^{(i)}$ be the distribution with p_i moved to the continuous part, *i.e.*,

$$P^{(i)} = (p_1, p_2, \ldots, p_{i-1}, p_{i+1}, \ldots, p_k)\,.$$

Using Expansion (2.4), we have

$$P(\bar{\psi}) = P_i(\bar{\psi}) + \varphi_1 p_i P_i(\bar{\psi}_m) + \sum_{t:\mu_t \geq 2} p_i^{\mu_t} P_i(\bar{\psi}_t).$$

On the other hand, using Expansion (2.7), we have

$$P^{(i)}(\bar{\psi}) = P_i(\bar{\psi}) + \varphi_1 p_i P_i(\bar{\psi}_m) + \sum_{\ell=2}^{\varphi_1} \binom{\varphi_1}{\ell} p_i^{\ell} P_i(\bar{\psi}_{[1\,..\,m-\ell]}).$$

Since P is the PML distribution, $P^{(i)}(\bar{\psi}) \leq P(\bar{\psi})$, and hence

$$\sum_{\ell=2}^{\varphi_1} \binom{\varphi_1}{\ell} p_i^{\ell} P_i(\bar{\psi}_{[1\,..\,m-\ell]}) \leq \sum_{t:\mu_t \geq 2} p_i^{\mu_t} P_i(\bar{\psi}_t).$$

Summing over all $i \in [k]$ yields

$$\sum_{\ell=2}^{\varphi_1} \binom{\varphi_1}{\ell} P(\bar{\psi}_{+\ell}) \leq \sum_{t:\mu_t \geq 2} P(\bar{\psi}), \tag{3.1}$$

where $\bar{\psi}_{+\ell}$ is the pattern obtained from $\bar{\psi}$ by identifying ℓ singletons, *i.e.*, $\bar{\psi}_{+\ell}$ has multiplicities

$$\mathcal{M}(\bar{\psi}_{+\ell}) = \{\mu_1, \mu_2, \ldots, \mu_{m-\ell}, \ell\}^* .$$

Note that

$$(\mu_1, \mu_2, \ldots, \mu_m) \preceq (\mu_1, \mu_2, \ldots, \mu_{m-\ell}, \ell, \underbrace{0, 0, ..., 0}_{k-m+\ell-1}).$$

By the majorization property in Fact 2.4, we have

$$P(\bar{\psi}) \leq (k - m + \ell - 1)^{\underline{\ell-1}} \cdot P(\bar{\psi}_{+\ell}).$$

Thus Equation (3.1) implies that

$$\sum_{\ell=2}^{\varphi_1} \frac{\binom{\varphi_1}{\ell}}{(k - m + \ell - 1)^{\underline{\ell-1}}} \leq m - \varphi_1.$$

Taking only the term for $\ell = 2$ on the left, we have

$$k \geq \frac{\binom{\varphi_1}{2}}{m - \varphi_1} + m - 1. \qquad \square$$

Theorem 3.3: Second Lower Bound

Proof of Theorem 3.3. For simplicity, let $P = (p_1, p_2, \ldots, p_k) = \hat{P}_{\bar{\psi}}$. For any $\{i, j\} \in \binom{[k]}{2}$, by Expansion (2.6),

$$P(\bar{\psi}) = P_{i,j}(\bar{\psi}) + \sum_t (p_i^{\mu_t} + p_j^{\mu_t}) P_{i,j}(\bar{\psi}_t) + \sum_{\{t,t'\} \in \binom{[m]}{2}} (p_i^{\mu_t} p_j^{\mu_{t'}} + p_i^{\mu_{t'}} p_j^{\mu_t}) P_{i,j}(\bar{\psi}_{t,t'}).$$

Let $P^{(i)}$ be the distribution with p_i split into two $\frac{p_i}{2}$'s. Using Expansion (2.6) again we have

$$\begin{aligned} P(\bar{\psi}) &= P_i(\bar{\psi}) + \sum_t p_i^{\mu_t} P_i(\bar{\psi}_t), \\ P^{(i)}(\bar{\psi}) &= P_i(\bar{\psi}) + \sum_t 2 \left(\frac{p_i}{2}\right)^{\mu_t} P_{i,j}(\bar{\psi}_t) + \sum_{\{t,t'\} \in \binom{[m]}{2}} 2 \left(\frac{p_i}{2}\right)^{\mu_t + \mu_{t'}} P_i(\bar{\psi}_{t,t'}). \end{aligned}$$

Since P is the PML distribution, $P^{(i)}(\bar{\psi}) \leq P(\bar{\psi})$, and hence

$$\sum_{t:\mu_t \geq 2} 2 \left(\frac{p_i}{2}\right)^{\mu_t} P_i(\bar{\psi}_t) + \sum_{\{t,t'\} \in \binom{[m]}{2}} 2 \left(\frac{p_i}{2}\right)^{\mu_t + \mu_{t'}} P_i(\bar{\psi}_{t,t'}) \leq \sum_{t:\mu_t \geq 2} p_i^{\mu_t} P_i(\bar{\psi}_t).$$

Summing over all $i \in [k]$,

$$\sum_{t:\mu_t \geq 2} \frac{1}{2^{\mu_t - 1}} P(\bar{\psi}) + \sum_{\{t,t'\} \in \binom{[m]}{2}} \frac{1}{2^{\mu_t + \mu_{t'} - 1}} P(\bar{\psi}_{+\{t,t'\}}) \leq (m - \varphi_1) P(\bar{\psi}),$$

where $\bar{\psi}_{+\{t,t'\}}$ is obtained from $\bar{\psi}$ by combining μ_t and $\mu_{t'}$, *i.e.*,

$$\mathcal{M}(\bar{\psi}_{+\{t,t'\}}) = \mathcal{M}(\bar{\psi}) \setminus \{\mu_t, \mu_{t'}\} \cup \{\mu_t + \mu_{t'}\}.$$

By the majorization property in Fact 2.4, we have

$$P(\bar{\psi}) \leq (k - m + 1) P(\bar{\psi}_{+\{t,t'\}}).$$

Then

$$\sum_{t:\mu_t \geq 2} \frac{1}{2^{\mu_t - 1}} + \sum_{t,t'} \frac{1}{2^{\mu_t + \mu_{t'} - 1}} \cdot \frac{1}{k - m + 1} \leq m - \varphi_1. \qquad \square$$

Theorem 3.5: Upper Bound on $\hat{k}$ for $\varphi_1 = 1$

Proof of Theorem 3.5. For simplicity, let $P = (p_1, p_2, \ldots, p_k) = \hat{P}_{\bar{\psi}}$. Let P' be the distribution with p_{k-1} and p_k merged, *i.e.*, replacing p_{k-1} and p_k by $p'_{k-1} = p_{k-1} + p_k$ and $p'_k = 0$. by Expansion (2.6),

$$P(\bar{\psi}) = P_{k-1,k}(\bar{\psi}) + \sum_t (p_{k-1}^{\mu_t} + p_k^{\mu_t}) P_{k-1,k}(\bar{\psi}_t) + \sum_{(t,t') \in [m]^{\underline{2}}} p_{k-1}^{\mu_t} p_k^{\mu_{t'}} P_{k-1,k}(\bar{\psi}_{t,t'}),$$

$$P'(\bar{\psi}) = P_{k-1,k}(\bar{\psi}) + \sum_t (p_{k-1} + p_k)^{\mu_t} P_{k-1,k}(\bar{\psi}_t).$$

Since P is the PML distribution, $P'(\bar{\psi}) \leq P(\bar{\psi})$. Then

$$\sum_{t:\mu_t \geq 2} \sum_{s=1}^{\mu_t - 1} \binom{\mu_t}{s} p_{k-1}^s p_k^{\mu_t - s} P_{k-1,k}(\bar{\psi}_t) \leq \sum_{(t,t') \in [m]^{\underline{2}}} p_{k-1}^{\mu_t} p_k^{\mu_{t'}} P_{k-1,k}(\bar{\psi}_{t,t'}). \tag{3.2}$$

Note that by Fact 2.6 P is discrete. Then, for any $(t, t') \in [m]^{\underline{2}}$,

$$P_{k-1,k}(\bar{\psi}_t) = \sum_{i=1}^{k-2} p_\ell^{\mu_{t'}} P_{k-1,k,i}(\bar{\psi}_{t,t'}) \geq p_{k-1}^{\mu_{t'}} \sum_{i=1}^{k-2} P_{k-1,k,i}(\bar{\psi}_{t,t'}),$$

where the equality holds only if P is uniform. Furthermore, since for each monomial term in $P_{k,k-1}(\bar{\psi}_{t,t'})$ there are $k - m$ missing indices $i \leq k - 2$, we have

$$\sum_{i=1}^{k-2} P_{k-1,k,i}(\bar{\psi}_{t,t'}) = (k - m) \sum_{i=1}^{k-2} P_{k-1,k}(\bar{\psi}_{t,t'}).$$

Thus

$$P_{k-1,k}(\bar{\psi}_t) \geq \frac{k-m}{m-1} \sum_{t' \neq t} p_{k-1}^{\mu_{t'}} P_{k-1,k}(\bar{\psi}_{t,t'}),$$

and hence the left-hand side LHS of Inequality (3.2) satisfies

$$\text{LHS} \geq \frac{k-m}{m-1} \sum_{t:\mu_t \geq 2} \sum_{s=1}^{\mu_t - 1} \binom{\mu_t}{s} \sum_{t' \neq t} p_{k-1}^{s} p_k^{\mu_t - s} \cdot p_{k-1}^{\mu_{t'}} P_{k-1,k}(\bar{\psi}_{t,t'}),$$

where the equality holds only if P is uniform. Note that

$$\sum_{s=1}^{\mu_t - 1} \binom{\mu_t}{s} p_{k-1}^{s} p_k^{\mu_t - s} \begin{cases} = 2p_{k-1}p_k = p_{k-1}p_k^{\mu_t - 1} + p_{k-1}^{\mu_t - 1}p_k, & \text{if } \mu_t = 2, \\ \geq \mu_t \left(p_{k-1}p_k^{\mu_t - 1} + p_{k-1}^{\mu_t - 1}p_k \right), & \text{if } \mu_t > 2, \end{cases}$$

where the equality holds only if $\mu_t = 3$. Thus

$$\text{LHS} \geq \frac{k-m}{m-1} \, \mu^*_{m-1} \sum_{(t,t'):\mu_t \geq 2} (p_{k-1}p_k^{\mu_t - 1} + p_{k-1}^{\mu_t - 1}p_k) \cdot p_{k-1}^{\mu_{t'}} P_{k-1,k}(\bar{\psi}_{t,t'}),$$

where the equality holds only if $\mu_t = 3$ for all $t \leq m-1$, and P is uniform. Combining with Inequality (3.2), we get

$$\frac{k-m}{m-1} \, \mu^*_{m-1} \sum_{(t,t'):\mu_t \geq 2} (p_{k-1}p_k^{\mu_t - 1} + p_{k-1}^{\mu_t - 1}p_k) \cdot p_{k-1}^{\mu_{t'}} P_{k-1,k}(\bar{\psi}_{t,t'}) \\ \leq \sum_{\{t,t'\} \in \binom{[m]}{2}} \left[(p_{k-1}^{\mu_t} p_k^{\mu_{t'}} + p_{k-1}^{\mu_{t'}} p_k^{\mu_t}) \right] P_{k-1,k}(\bar{\psi}_{t,t'}).$$

Note that the left-hand side has more terms (strictly more if $m \geq 3$), and on the right-hand side either $\mu_t \geq 2$ or $\mu_{t'} \geq 2$ since $\varphi_1 = 1$. Further notice that

$$(p_{k-1}p_k^{\mu_t - 1} + p_{k-1}^{\mu_t - 1}p_k) \cdot p_{k-1}^{\mu_{t'}} \geq (p_{k-1}^{\mu_t} p_k^{\mu_{t'}} + p_{k-1}^{\mu_{t'}} p_k^{\mu_t}).$$

It follows that $\frac{k-m}{m-1} \, \mu^*_{m-1} \leq 1$, *i.e.*,

$$k \leq m + \frac{m-1}{\mu^*_{m-1}},$$

where the equality holds only if P is uniform and $\bar{\psi} = 1112$. However, it was shown in Fact 2.12 that $\hat{P}_{1112} = \left(\frac{1}{2}, \frac{1}{2}\right)$. Thus the strict inequality holds for all patterns with $\varphi_1 = 1$. $\square$

Theorem 3.4: Upper Bound on $\hat{k}$ for Large φ_2

Proof of Theorem 3.4. For simplicity, let $P = (p_1, p_2, \ldots, p_k) = \hat{P}_{\bar{\psi}}$. Without loss of generality, assume that $p_k > \frac{1}{n-1}$. For any $i \in [k]$, by Expansion (2.4),

$$P(\bar{\psi}) = P_i(\bar{\psi}) + \varphi_1 p_i P_i(\bar{\psi}_m) + \sum_{t:\mu_t \geq 2} p_i^{\mu_t} P_i(\bar{\psi}_t).$$

Note that

$$P_i(\bar{\psi}) \leq (1-p_i)^n \hat{P}(\bar{\psi}) = (1-p_i)^n P(\bar{\psi}) \leq \left[1 - np_i + \binom{n}{2} p_i^2\right] P(\bar{\psi}).$$

Then

$$P(\bar{\psi}) \leq \left[1 - np_i + \binom{n}{2} p_i^2\right] P(\bar{\psi}) + \varphi_1 p_i P_i(\bar{\psi}_m) + \sum_{t:\mu_t \geq 2} p_i^{\mu_t} P_i(\bar{\psi}_t),$$

i.e.,

$$\left(1 - \frac{n-1}{2} p_i\right) \cdot nP(\bar{\psi}) \leq \varphi_1 P_i(\bar{\psi}_m) + \sum_{t:\mu_t \geq 2} p_i^{\mu_t - 1} P_i(\bar{\psi}_t).$$

On the other hand,

$$nP(\bar{\psi}) = \frac{\partial P(\bar{\psi})}{\partial p_i} = \varphi_1 P_i(\bar{\psi}_m) + \sum_{t:\mu_t \geq 2} \mu_t p_i^{\mu_t - 1} P_i(\bar{\psi}_t).$$

Then

$$\left[\left(1 - \frac{n-1}{2} p_i\right) \cdot 2 - 1\right] \sum_{t:\mu_t \geq 2} p_i^{\mu_t - 1} P_i(\bar{\psi}_t) \leq \frac{n-1}{2} p_i \cdot \varphi_1 P_i(\bar{\psi}_m).$$

For $\varphi_2 > 0$,

$$(1 - (n-1)p_i) \cdot \varphi_2 P_i(\bar{\psi}_1) \leq \frac{n-1}{2} \cdot \varphi_1 P_i(\bar{\psi}_m).$$

Since

$$P_i(\bar{\psi}_1) \geq \frac{m-1}{1-p_i} P_i(\bar{\psi}_m),$$

we have

$$(1 - (n-1)p_i) \cdot \varphi_2 (m-1) \leq \frac{n-1}{2} \cdot \varphi_1 (1-p_i).$$

By assumption, $\varphi_2 \cdot 2(m-1) \geq c(n-1)\varphi_1$. Then

$$(1 - (n-1)p_i) \cdot c \leq 1 - p_i.$$

Solving for p_i,

$$p_i \geq \frac{c-1}{(n-1)c-1} \geq \frac{c-1}{cn}.$$

Then

$$k \leq \frac{(n-1)c-1}{c-1} < \frac{c}{c-1}(n-1). \qquad \square$$

Theorem 3.6: Implicit Upper Bound on $\hat{k}$ for $\varphi_2 = 0$

Proof of Theorem 3.6. For simplicity, let $P = (p_1, p_2, \ldots, p_k) = \hat{P}_{\bar{\psi}}$. For any $i \in [k]$, Similar to Equation (3.1) in the proof of Theorem 3.2, we can show that

$$\sum_{\ell=2}^{\varphi_1} \binom{\varphi_1}{\ell} p_i^{\ell} P_i(\bar{\psi}_{[1\,..\,m-\ell]}) \leq \sum_{t:\mu_t \geq 2} p_i^{\mu_t} P_i(\bar{\psi}_t).$$

Note that, for any $\ell \geq 2$,

$$P_i(\bar{\psi}_m) \leq (1-p_i)^{\ell-1} P_i(\bar{\psi}_{[1\,..\,m-\ell]}).$$

Taking only the term for $\ell = 2$, we get

$$\binom{\varphi_1}{2} \frac{p_i^2}{1-p_i} P_i(\bar{\psi}_m) \leq \sum_{t:\mu_t \geq 2} p_i^{\mu_t} P_i(\bar{\psi}_t). \tag{3.3}$$

Since $P_i(\bar{\psi}) \leq (1-p_i) P_i(\bar{\psi}_m)$,

$$\begin{aligned} P(\bar{\psi}) &= P_k(\bar{\psi}) + \varphi_1 p_k P_k(\bar{\psi}_m) + \sum_{t:\mu_t \geq 2} p_k^{\mu_t} P_k(\bar{\psi}_t) \\ &\leq (1 + (\varphi_1 - 1)p_k)\, P_k(\bar{\psi}_m) + \sum_{t:\mu_t \geq 2} p_k^{\mu_t} P_k(\bar{\psi}_t). \end{aligned} \tag{3.4}$$

Canceling $P_k(\bar{\psi}_m)$ from Inequalities (3.3) and (3.4), we have

$$\binom{\varphi_1}{2} p_i^2 P(\bar{\psi}) \leq \left[\binom{\varphi_1 - 1}{2} p_i^2 + (\varphi_1 - 2)p_i + 1 \right] \sum_{t:\mu_t \geq 2} p_i^{\mu_t} P_i(\bar{\psi}_t) \tag{3.5}$$

Next we show that

Claim 3.1. $\sum_{t:\mu_t \geq 2} p_k^{\mu_t} P_k(\bar{\psi}_t) \leq (m - \varphi_1) p_k^{\mu_*} P_k(\bar{\psi}_*)$.

To this see, note that for each $t \in [m]$ such that $\mu_t \geq 2$ and $t \neq m - \varphi_1$,

$$\begin{aligned} p_k^{\mu_t} P_k(\bar{\psi}_t) - p_k^{\mu_*} P_k(\bar{\psi}_*) &= p_k^{\mu_t} \sum_{i \neq k} p_k^{\mu_*} P_{k,i}(\bar{\psi}_{t,m-\varphi_1}) - p_k^{\mu_*} \sum_{i \neq k} p_k^{\mu_t} P_{k,i}(\bar{\psi}_{t,m-\varphi_1}) \\ &= \sum_{i \neq k} P_{k,i}(\bar{\psi}_{t,m-\varphi_1})(p_k^{\mu_t} p_k^{\mu_*} - p_k^{\mu_*} p_k^{\mu_t}) \leq 0. \end{aligned}$$

Thus Claim 3.1 holds. Further note that

$$P_k(\bar{\psi}_*) \le (1-p_k)^{n-\mu_*}\hat{P}(\bar{\psi}_*) \le (1-p_k)^{\varphi_1}\hat{P}(\bar{\psi}_*).$$

Then

$$\sum_{t:\mu_t\ge 2} p_k^{\mu_t}P_k(\bar{\psi}_t) \le (m-\varphi_1)p_k^{\mu_*} \le (m-\varphi_1)p_k^{\mu_*-2}(1-p_k)^{\varphi_1}\hat{P}(\bar{\psi}_*).$$

Combining with Inequality (3.5), we get

$$\binom{\varphi_1}{2}\hat{P}(\bar{\psi}) \le \left[\binom{\varphi_1-1}{2}p_k^2 + (\varphi_1-2)p_k + 1\right](m-\varphi_1)p_k^{\mu_*-2}(1-p_k)^{\varphi_1}\hat{P}(\bar{\psi}_*).$$

It can verified that

$$\left[\binom{\varphi_1-1}{2}p_k^2 + (\varphi_1-2)p_k + 1\right](1-p_k)^{\varphi_1} \le 1$$

by showing that the left-hand side decreases in $p_k \in [0,1]$. Therefore

$$(m-\varphi_1)p_k^{\mu_*-2} \ge \binom{\varphi_1}{2}\frac{\hat{P}(\bar{\psi})}{\hat{P}(\bar{\psi}_*)}.$$

Solving for p_k completes the proof. □

Corollary 3.7: Explicit Upper Bound on $\hat{k}$ for $\varphi_2 = 0$

Proof of Corlollary 3.7. In general, by combinatorial arguments it is easy to see that the number of patterns having the same canonical pattern $1^{\mu_1}2^{\mu_2}\cdots m^{\mu_m}$ is the number of partitions of the set $[n]$ into disjoint subsets of sizes $\mu_1, \mu_2, \ldots, \mu_m$, *i.e.*,

$$\frac{\binom{n}{\mu_1,\mu_2,\ldots,\mu_m}}{\prod_{\mu>0}\varphi_\mu!} = \frac{n!/\prod_{t=1}^m \mu_t!}{\prod_{\mu>0}\varphi_\mu} = \frac{n!}{\prod_{t=1}^m \mu_t!\prod_{\mu>0}\varphi_\mu!}.$$

Thus

$$\hat{P}(\bar{\psi}) \le \frac{\prod_{t=1}^m \mu_t!\prod_{\mu>0}\varphi_\mu!}{n!}.$$

Here we are interested in an upper bound for $\hat{P}(\bar{\psi}_r)$. For simplicity, let $r = m-\varphi_1$. Then $\mu_* = \mu_r$ and $\bar{\psi}_* = \bar{\psi}_r$, and

$$\hat{P}(\bar{\psi}_r) \le \frac{\prod_{t=1}^r \mu_t!\prod_{\mu>0}\varphi_\mu!}{(n-\mu_r)!\cdot\mu_r!\varphi_{\mu_r}}.$$

On the other hand, using distribution $P' = \left(\frac{\mu_1}{n}, \frac{\mu_2}{n}, \ldots, \frac{\mu_r}{n}\right)$ we get a lower bound on $\hat{P}(\bar{\psi})$:

$$\hat{P}(\bar{\psi}) \geq \prod_{t=1}^{r} \left(\frac{\mu_t}{n}\right)^{\mu_t} \prod_{\mu>1} \varphi_\mu! \cdot \left(\frac{\varphi_1}{n}\right)^{\varphi_1}.$$

It follows that

$$\frac{\hat{P}(\bar{\psi}_r)}{\hat{P}(\bar{\psi})} \leq \frac{1}{\varphi_{\mu_r}} \cdot \frac{\prod_{t=1}^{r} \mu_t!}{\prod_{t=1}^{r} \mu_t^{\mu_t}} \cdot \frac{n^n}{(n-\mu_r)!\mu_r!} \cdot \frac{\varphi_1!}{\varphi_1^{\varphi_1}}.$$

Note that

$$\frac{n^n}{(n-\mu_r)!\mu_r!} = \frac{n^n}{n!}\binom{n}{\mu_r} \leq \frac{n^n}{n!}\binom{n}{n/2} = \frac{n^n}{(n/2)!(n/2)!}.$$

Furthermore, for any integer $n > 0$ from [Rob55] the factorial $n!$ can be bounded as

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}.$$

Thus

$$\frac{\hat{P}(\bar{\psi}_r)}{\hat{P}(\bar{\psi})} \leq \frac{1}{\varphi_{\mu_r}} \cdot \prod_{t=1}^{r} \frac{\sqrt{2\pi\mu_t}\, e^{\frac{1}{12\mu_t}}}{e^{\mu_t}} \cdot \frac{n^n\, e^n}{\pi n\, (n/2)^n\, e^{\frac{2}{6n+1}}} \cdot \frac{\sqrt{2\pi\varphi_1} e^{\frac{1}{12\varphi_1}}}{e^{\varphi_1}}.$$

Note that

$$\prod_{t=1}^{r} \sqrt{2\pi\mu_t} \cdot \sqrt{2\pi\varphi_1} \leq \left(\frac{2\pi n}{r+1}\right)^{\frac{r+1}{2}} \leq e^{\frac{\pi n}{e}},$$

and

$$n = \sum_{t=1}^{r} \mu_t + \varphi_1 > 3r.$$

Then

$$\frac{\hat{P}(\bar{\psi}_r)}{\hat{P}(\bar{\psi})} \leq \frac{1}{\varphi_{\mu_r}} \cdot e^{\frac{\pi n}{e}} \cdot \frac{2^n}{\pi n} \cdot e^{\sum_{t=1}^{r} \frac{1}{12\mu_t} + \frac{1}{12\varphi_1} - \frac{2}{6n+1}} < e^{\frac{\pi n}{e} + n\ln 2 + \frac{n/3}{12\cdot 3}} < e^{2n}.$$

The conclusion follows from Theorem 3.6. □

Theorem 3.8: Another Bound on $\hat{k}$

Proof of Theorem 3.8. For simplicity, let $P = \hat{P}_{\bar{\psi}}$ and let $p_1 > p_2 > \cdots > p_d > 0$ be the distinct probabilities. Furthermore, let $k_i = \nu(p_i)$ and $\alpha_i = k_i p_i$. For any $S \subseteq [m]$, let

$$m_S \stackrel{\text{def}}{=} |S| \text{ and } n_S \stackrel{\text{def}}{=} \sum_{t \in S} \mu_t.$$

Then, for a given $i \in [d]$,

$$P(\bar{\psi}) = \sum_{S \subseteq [m]} \left(\frac{1}{k_i}\right)^{n_S - m_S} \cdot \frac{k_i^{m_S}}{k_i^{m_S}} \alpha_i^{n_S} P_i(\bar{\psi}_{\bar{S}}) = \sum_{S \subseteq [m]} f_S(x) \cdot \alpha_i^{n_S} P_i(\bar{\psi}_{\bar{S}}),$$

where we used $x = \frac{1}{k_i}$ for simplicity, and

$$f_S(x) \stackrel{\text{def}}{=} x^{n_S - m_S} \prod_{t=1}^{m_S - 1} (1 - tx).$$

We consider the monotonicity of $f_S(x)$ for $x \leq \frac{1}{m}$. By direct calculation, we have

$$f_S'(x) = x^{n_S - m_S - 1} \prod_{t=1}^{m_S - 1} (1 - tx) \cdot \left[n_S - 1 - \sum_{t=1}^{m_S - 1} \frac{1}{1 - tx} \right],$$

$$f_S''(x) = x^{n_S - m_S - 2} \prod_{t=1}^{m_S - 1} (1 - tx) \cdot \left[\left(n_S - 1 - \sum_{t=1}^{m_S - 1} \frac{1}{1 - tx} \right) \left(n_S - 2 - \sum_{t=1}^{m_S - 1} \frac{1}{1 - tx} \right) - \sum_{t=1}^{m_S - 1} \frac{tx}{(1 - tx)^2} \right].$$

Case 1 $\varphi_1 = 0$. Suppose $x < \frac{n_S - m_S}{(m_S - 1)(n_S - 1)}$. Then, for all $S \subseteq [m]$ such that $S \neq \emptyset$,

$$n_S - 1 - \sum_{t=1}^{m_S - 1} \frac{1}{1 - tx} \geq n_S - 1 - \frac{m_S - 1}{1 - (m_S - 1) \cdot \frac{n_S - m_S}{(m_S - 1)(n_S - 1)}} = 0.$$

Then $f_S'(x) \geq 0$, where the strict inequality holds if $m_S \geq 2$ or $n_S \geq 2$, which is true since $\varphi_1 = 0$ and $S \neq \emptyset$. In other words, $P(\bar{\psi})$ increases in k_i if, for all $S \subseteq [m]$ such that $S \neq \emptyset$,

$$k_i > \frac{(m_S - 1)(n_S - 1)}{n_S - m_S}.$$

Since the value of k_i defined for $P = \hat{P}_{\bar{\psi}}$ is optimal, we must have

$$f_S\left((k_i \pm 1)^{-1}\right) \leq f_S(k_i^{-1}),$$

which means that $f_S(x)$ is not strictly increasing for $x \leq \frac{1}{k_i - 1}$; otherwise $k_i - 1$ gives a higher pattern probability. Thus

$$k_i - 1 \leq \max_{S \subseteq [m]} \frac{(m_S - 1)(n_S - 1)}{n_S - m_S} \leq \max_{S \subseteq [m]} \frac{(m_S - 1)(2m_S - 1)}{m_S} \leq 2m - 1 + \frac{1}{m},$$

where the second inequality follows from the fact that $\varphi_1 = 0$ and hence $n_S \geq 2m_S$.

Case 2 $\varphi_2 > 0$. We look for a sufficient condition so that $f_S''(x) > 0$ for all $S \neq \emptyset$. Note that since $\varphi_2 = 0$, for any $S \subseteq [m]$ either $n_S = m_S$ or $n_S - m_S \geq 2$.

- If $n_S - m_S = 0$, then

$$f_S''(x) = \prod_{t=1}^{m_S-1} (1-tx) \cdot \sum_{t_1 \neq t_2} \frac{t_1 t_2}{(1-t_1 x)(1-t_2 x)} > 0.$$

- If $n_S - m_S \geq 2$, then $f_S''(x) > 0$ if

$$\left(n_S - 1 - \sum_{t=1}^{m_S-1} \frac{1}{1-tx}\right)\left(n_S - 2 - \sum_{t=1}^{m_S-1} \frac{1}{1-tx}\right) > \sum_{t=1}^{m_S-1} \frac{tx}{(1-tx)^2},$$

which is clearly true if $m_S = 1$. Without loss of generality we assume $m_S \geq 2$. Note that

$$\sum_{t=1}^{m_S-1} \frac{tx}{(1-tx)^2} \leq \frac{(m_S-1)^2 x}{[1-(m_S-1)x]^2} = \frac{(m_S-1)y}{(1-y)^2},$$
$$\sum_{t=1}^{m_S-1} \frac{1}{1-tx} \leq \frac{m_S-1}{1-(m_S-1)x} = \frac{m_S-1}{1-y},$$

where

$$y = (m_S - 1)x.$$

Then $f_S''(x) > 0$ if

$$n_S - 2 - \frac{m_S-1}{1-y} \geq 0, \tag{3.6}$$
$$\left(n_S - 1 - \frac{m_S-1}{1-y}\right)\left(n_S - 2 - \frac{m_S-1}{1-y}\right) > \frac{(m_S-1)y}{(1-y)^2}. \tag{3.7}$$

Assume that $x \leq \frac{1}{m}$. Then $y < 1$. Solving Inequality (3.6) for $y < 1$, we get

$$y \leq \frac{n_S - m_S - 1}{n_S - 1},$$

or equivalently

$$x \leq \frac{n_S - m_S - 1}{(m_S-1)(n_S-1)}.$$

Inequality (3.7) can be written as

$$(n_S-1)(n_S-2)y^2 - 2(n_S-m_S)(n_S-2)y + (n_S-m_S)(n_S-m_S-1) > 0. \tag{3.8}$$

Since we assumed that $n_S - m_S \geq 2$, we have $n_S \geq 3$. Then

$$y \leq \frac{n_S - m_S - 1}{n_S - 1} \leq \frac{1}{2}(n_S - m_S - 1),$$

and hence

$$(n_S - 1)(n_S - 2)y^2 > 0,$$
$$-2(n_S - m_S)(n_S - 2)y + (n_S - m_S)(n_S - m_S - 1) \geq 0.$$

It follows that Inequality (3.8) holds, and hence Inequality (3.7) holds.

In summary, $P(\bar{\psi})$ is strictly convex in $x = \frac{1}{k_i}$ if $x \leq \frac{n_S - m_S - 1}{(m_S - 1)(n_S - 1)}$ for all $S \subseteq [m]$ such that $n_S - m_S \geq 2$. Support $k_i > m$, then we must have

$$f_S((k_i \pm 1)^{-1}) \leq f_S(k_i^{-1}),$$

which means that $f_S(x)$ is not strictly convex for $x \leq \frac{1}{k_i - 1}$; otherwise either $k_i - 1$ or $k_i + 1$ gives a higher pattern probability. It follows that

$$\frac{1}{k_i - 1} > \min_{S \subseteq [m]: n_S - m_S \geq 2} \frac{n_S - m_S - 1}{(m_S - 1)(n_S - 1)},$$

i.e.,

$$k_i < \max_{S \subseteq [m]: n_S - m_S \geq 2} \frac{m_S(n_S - 2)}{n_S - m_S - 1} \leq m^2.$$

From Fact 2.7 we know that $d \leq \min\{n - 1, 2^m\}$. Thus

$$\hat{k} \leq \max_{i \in [d]} k_i \cdot \min\{n - 1, 2^m\} \leq m^2 \cdot \min\{n - 1, 2^m\}. \qquad \square$$

3.3 Continuous Probability

We given a sufficient condition such that the PML distribution has positive continuous probability, as well as a sufficient condition such that it has no continuous part.

3.3.1 Patterns with $\hat{q} > 0$

Recall that μ_* is the smallest multiplicity greater than 1, *i.e.*$\mu_* = \mu_{m-\varphi_1}$, and $\bar{\psi}_* \stackrel{\text{def}}{=} \bar{\psi}_{m-\varphi_1}$.

Theorem 3.9. *For any pattern such that* $\varphi_2 = 0$ *and* $e^{\frac{2n}{\mu_*-1}} \leq \frac{\binom{\varphi_1}{2}}{m-\varphi_1}$, *the PML distribution satisfies*

$$\hat{k} \leq \frac{\binom{\varphi_1}{2}}{m-\varphi_1} \text{ and } \hat{q} > 0.$$

Proof. By Corollary 3.7

$$\hat{k} < \sqrt[\mu_*-2]{\frac{m-\varphi_1}{\binom{\varphi_1}{2}} \cdot e^{2n}} \leq \frac{\binom{\varphi_1}{2}}{m-\varphi_1}.$$

This proves the first part. For the second part, suppor $\hat{q} = 0$. Then it follows from Theorem 3.2 that

$$\hat{k} \geq \frac{\binom{\varphi_1}{2}}{m-\varphi_1} + m - 1 > \frac{\binom{\varphi_1}{2}}{m-\varphi_1},$$

a contradiction. Thus we must have $\hat{q} > 0$. □

3.3.2 Patterns with $\hat{q} = 0$

It was previously shown that, as shown in Fact 2.6, for any pattern $\bar{\psi} \neq 1$ such that $\varphi_1 = 0$ or $\varphi_1 = 1$, th PML distribution is discrete. In this section we extend the results by providing a more general condition.

Theorem 3.10. *For any pattern* $\bar{\psi}$, *if*

$$\varphi_2 \geq \frac{n-1}{2(m-1)}\binom{\varphi_1}{2},$$

then $\hat{q} = 0$.

Proof. For simplicity, let $P = (p_1, p_2, \ldots, p_k) = \hat{P}_{\bar{\psi}}$. Suppose $q > 0$. For any $\varepsilon \in [0, q]$ let P^ε be the distribution obtained from P with continuous probability

ε, out of q, moved to a new discrete symbol. Then

$$P(\bar{\psi}) = P_\varepsilon(\bar{\psi}) + \sum_{\ell \geq 1} \binom{\varphi_1}{\ell} \varepsilon^\ell P_\varepsilon(\bar{\psi}_{[1\,..\,m-\ell]}),$$

and

$$Q^\varepsilon(\bar{\psi}) = P_\varepsilon(\bar{\psi}) + \sum_{t=1}^{m} \varepsilon^{\mu_t} P_\varepsilon(\bar{\psi}_t),$$

Let $r = m - \varphi_1$. Since P is the PML distribution, $P^\varepsilon(\bar{\psi}) \leq P(\bar{\psi})$, and hence

$$\varphi_2 \varepsilon^2 P_\varepsilon(\bar{\psi}_r) + \sum_{t:\mu_t \geq 3} \varepsilon^{\mu_t} P_\varepsilon(\bar{\psi}_t) \leq \binom{\varphi_1}{2} \varepsilon^2 P_\varepsilon(\bar{\psi}_{m-1,m}) + \sum_{\ell \geq 3} \binom{\varphi_1}{\ell} \varepsilon^\ell P_\varepsilon(\bar{\psi}_{[1\,..\,m-\ell]}),$$

i.e.,

$$\varphi_2 P_\varepsilon(\bar{\psi}_r) + \sum_{t:\mu_t \geq 3} \varepsilon^{\mu_t - 2} P_\varepsilon(\bar{\psi}_t) \leq \binom{\varphi_1}{2} P_\varepsilon(\bar{\psi}_{m-1,m}) + \sum_{\ell \geq 3} \binom{\varphi_1}{\ell} \varepsilon^{\ell-2} P_\varepsilon(\bar{\psi}_{[\ell]}),$$

Letting $\varepsilon \to 0$ we get

$$\varphi_2 P(\bar{\psi}_r) \leq \binom{\varphi_1}{2} P(\bar{\psi}_{m-1,m}).$$

Note that

$$\begin{aligned} P(\bar{\psi}_r) &= \sum_{t \neq r} P(\bar{\psi}_{t,r} \otimes 1^{\mu_t+1}) + P(\bar{\psi}_r \otimes 1) \\ &> (m-1) P(\bar{\psi}_m) + P(\bar{\psi}_r) \left(1 - \sum_{t=1}^{m-1} \frac{\mu_t}{n}\right) \\ &\geq (m-1) P(\bar{\psi}_m) + P(\bar{\psi}_r) \cdot \frac{1}{n}, \end{aligned}$$

where the first inequality follows from (i) $\bar{\psi}_{r,t} \otimes 1^{\mu_t+1} \succeq \bar{\psi}_m$ and Fact 2.4, and (ii) Fact 2.3. Note that the inequality is strict since $q > 0$. Then

$$P(\bar{\psi}_r) \geq \frac{n(m-1)}{n-1} P(\bar{\psi}_m).$$

Moreover,

$$P(\bar{\psi}_m) \geq P(\bar{\psi}_{m-1,m}) \cdot \left(1 - \sum_{t=1}^{m-2} \frac{\mu_t}{n}\right) \geq P(\bar{\psi}_{m-1,m}) \cdot \frac{2}{n}.$$

It follows that

$$(m-1)\varphi_2 \leq \binom{\varphi_1}{2} \frac{P(\bar{\psi}_{m-1,m})}{P(\bar{\psi}_m)} < \binom{\varphi_1}{2} \cdot \frac{n-1}{2},$$

a contradiction. □

Chapter 4

Patterns of Simple Forms

Chapter 2 analytically determines the PML distributions for trival, uniform, 1-uniform, binary and some of the skewed patterns. In this chapter we find the PML distributions for patterns of more general forms.

- In Section 4.1 we give an alternative proof for binary patterns.
- In Section 4.2 we prove that Conjecture 1 holds for all truly skewed patterns.
- In Section 4.3 we show that quasi-uniform patterns have uniform PML.
- In Section 4.4 we further extend quasi-uniform to almost-uniform patterns.

4.1 Binary Patterns Revisited

As stated in Fact 2.12, the PML distribution of a non-trivial binary pattern is $\left(\frac{1}{2}, \frac{1}{2}\right)$ if $(\mu_1 - \mu_2)^2 \leq n$. We give a proof, different from the original proof, using induction and the Inequality of Arithmetic and Geometric Means (AM-GM Inequality).

Theorem 4.1. *For any $p \in [0,1]$, $q = 1 - p$, and integers $a \geq 0, b \geq 0$ such that $p + q = 1$ and $(a-b)^2 \leq a + b$,*

$$p^a q^b + p^b q^a \leq 2\left(\frac{1}{2}\right)^{a+b},$$

where the equality holds if and only if $p = q = \frac{1}{2}$.

Proof. Without loss of generality, assume $a \geq b$. Let $\delta \stackrel{\text{def}}{=} a - b > 0$. It is easy to verify that the condition $(a-b)^2 \leq a + b$ is equivalent to

$$b \geq \binom{\delta}{2},$$

and

$$p^a q^b + p^b q^a = (pq)^b (p^\delta + q^\delta).$$

We use induction on $a + b = 2b + \delta$. For $2b + \delta = 3$, *i.e.*, $a = 2, b = 1$,

$$p^a q^b + p^b q^a = p^2 q + pq^2 = pq \leq \frac{1}{4} = 2\left(\frac{1}{2}\right)^{2+1}.$$

For $2b + \delta > 3$, suppose $(pq)^{b'}(p^{\delta'} + q^{\delta'}) \leq 2\left(\frac{1}{2}\right)^{2b'+\delta'}$ for all $b' \geq 0, \delta' \geq 0$ such that $2b' + \delta' < 2b + \delta$. We consider two cases.

- If δ is even, let $\delta = 2k$. Then

$$(pq)^b (p^\delta + q^\delta) = \frac{1}{2}(pq)^{b-k} \cdot 2(pq)^k \left[(p^k + q^k)^2 - 2(pq)^k\right].$$

  By AM-GM Inequality,

$$\begin{aligned} 2(pq)^k \left[(p^k + q^k)^2 - 2(pq)^k\right] &\leq \frac{1}{4}\left[2(pq)^k + (p^k + q^k)^2 - 2(pq)^k\right]^2 \\ &= \frac{1}{4}(p^k + q^k)^4. \end{aligned}$$

  Then

$$(pq)^b (p^\delta + q^\delta) \leq \frac{1}{2}(pq)^{b-k} \cdot \frac{1}{4}(p^k + q^k)^4 = \frac{1}{8}(pq)^{b-k-4\binom{k}{2}} \left[(pq)^{\binom{k}{2}}(p^k + q^k)\right]^4.$$

  Let $\delta' = k, b' = \binom{k}{2}$. By inductive hypothesis,

$$(pq)^{\binom{k}{2}}(p^k + q^k) = (pq)^{b'}(p^{\delta'} + q^{\delta'}) \leq 2\left(\frac{1}{2}\right)^{2b'+\delta'} = 2\left(\frac{1}{2}\right)^{k^2}.$$

  Furthermore, it is easy to see that $pq \leq \left(\frac{1}{2}\right)^2$. Thus

$$(pq)^b (p^\delta + q^\delta) \leq \frac{1}{8}\left(\frac{1}{2}\right)^{2b-2k-8\binom{k}{2}} \cdot \left[2\left(\frac{1}{2}\right)^{k^2}\right]^4 = 2\left(\frac{1}{2}\right)^{2b+2k}.$$

- If δ is odd, let $\delta = 2k+1$. Then

$$(pq)^b(p^\delta + q^\delta) = (pq)^b\left[(p^k+q^k)(p^{k+1}+q^{k+1}) - (pq)^k\right].$$

By AM-GM Inequality,

$$(pq)^k\left[(p^k+q^k)(p^{k+1}+q^{k+1}) - (pq)^k\right] \le \frac{1}{4}\left[(p^k+q^k)(p^{k+1}+q^{k+1})\right]^2.$$

Then

$$\begin{aligned}(pq)^b(p^\delta+q^\delta) &\le \frac{1}{4}(pq)^{b-k}\left[(p^k+q^k)(p^{k+1}+q^{k+1})\right]^2 \\ &= \frac{1}{4}(pq)^{b-k-2\binom{k}{2}-2\binom{k+1}{2}} \\ &\quad\cdot\left[(pq)^{\binom{k}{2}}(p^k+q^k)(pq)^{\binom{k+1}{2}}(p^{k+1}+q^{k+1})\right]^2.\end{aligned}$$

By inductive hypothesis, we have

$$\begin{aligned}(pq)^{\binom{k}{2}}(p^k+q^k) &\le 2\left(\frac{1}{2}\right)^{2\binom{k}{2}+k}, \\ (pq)^{\binom{k+1}{2}}(p^{k+1}+q^{k+1}) &\le 2\left(\frac{1}{2}\right)^{2\binom{k+1}{2}+k+1}.\end{aligned}$$

Thus

$$(pq)^b(p^\delta+q^\delta) \le \frac{1}{4}\left(\frac{1}{2}\right)^{2b-2k-4\binom{k}{2}-4\binom{k+1}{2}}\left[2\left(\frac{1}{2}\right)^{k^2}\cdot 2\left(\frac{1}{2}\right)^{(k+1)^2}\right]^2,$$

which can be simplified to $2\left(\frac{1}{2}\right)^{2b+2k+1}$. □

4.2 Skewed Patterns

Recall that a truly skewed pattern has the form $\bar{\psi} = 1^r23\cdots(u+1)$, where $r \ge 3$ and $u \ge 2$. As stated in Conjecture 1, it was believed that $\hat{P}_{1^r23\cdots(u+1)} = \left(\frac{r}{r+u}\right)$, which has been shown to hold for $r \ge 2\sqrt{u} \gg 1$. In this section we show that the conjecture indeed holds for all truly skewed patterns:

Theorem 4.2. *For any $r \ge 3$ and $u \ge 2$,*

$$\hat{P}_{1^r23\cdots(u+1)} = \left(\frac{r}{r+u}\right).$$

The structure of the remaining part of this section is illustrated in Figure 4.1.

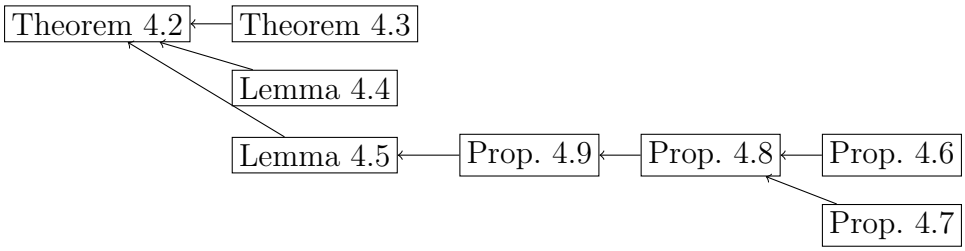


Figure 4.1: Roadmap to the Proof for Skewed Patterns

4.2.1 Pattern 11123

We first give a proof to the smallest truly skewed pattern $\bar{\psi}_{3,2} = 11123$ and then extend it to the others.

Theorem 4.3.

$$\hat{P}_{11123} = \left(\frac{3}{5}\right).$$

Proof. For simplicity, let $P = (p_1, p_2, \ldots, p_k) = \hat{P}_{11123} \in \mathcal{P}_{\text{mix}}^{\text{sorted}}$. We first show that p_1 can't be too small:

Claim 4.1. $p_1 > 0.4549$.

To show Claim 4.1, by Lemma 3.1 we have

$$5p_iP(11123) = 3p_i^3P_i(12) + 2p_iP_i(1112). \tag{4.1}$$

On the other hand, expanding $P(11123)$ by p_i, we have

$$P(11123) = P_i(11123) + p_i^3P_i(12) + 2p_iP_i(1112). \tag{4.2}$$

Using $3 \times (4.2) - (4.1)$, we get

$$\begin{aligned}(3 - 5p_i)P(11123) &= 3P_i(11123) + 4p_iP_i(1112) \\ &\leq 3(1-p_i)^5P(11123) + 4p_i(1-p_i)^4 \cdot \frac{1}{8},\end{aligned}$$

i.e.,

$$\left[(3 - 5p_i) - 3(1-p_i)^5\right] P(11123) \leq \frac{1}{2}p_i(1-p_i)^4.$$

Since P is the PML distribution, $P(11123) \geq \left(\frac{3}{5}\right)^3 \left(\frac{2}{5}\right)^2$, and hence

$$\left[(3-5p_i) - 3(1-p_i)^5\right] \left(\frac{3}{5}\right)^3 \left(\frac{2}{5}\right)^2 \leq \frac{1}{2} p_i (1-p_i)^4.$$

Solving for p_i gives

$$p_i < 0.3531, \text{ or } p_i > 0.4549.$$

However,

$$\left(\frac{3}{5}\right)^3 \left(\frac{2}{5}\right)^2 \leq P(11123) \leq p_1 P(1123) \leq p_1 \left(\frac{1}{5}\right)^4 \cdot (5 \cdot 4 \cdot 3),$$

which implies $p_1 \geq 0.36$. Thus we must have $p_1 > 0.4549$.

Next we show that

Claim 4.2. *For all $i \in [k]$,*

$$10p_i^3 - 8p_i^2 + 3p_i \geq 3p_i P(11) + 2P(111).$$

To show Claim 4.2, we rewrite $P(11123)$ as

$$\begin{aligned}
P(11123) &= P(1112) - [P(11112) + P(11122)] \\
&= [P(111) - P(1111)] - [P(1111) - P(11111)] \\
&\quad - [P(111)P(11) - P(11111)] \\
&= P(111) - 2P(1111) - P(111)P(11) + 2P(11111) \\
&= f(p_1, p_2, \ldots, p_k) \stackrel{\text{def}}{=} \sum_{i=1}^{k} p_i^3 - 2\sum_{i=1}^{k} p_i^4 - \sum_{i=1}^{k} p_i^3 \sum_{i=1}^{k} p_i^2 + 2\sum_{i=1}^{k} p_i^5.
\end{aligned}$$

Note that $q = 1 - \sum_{i=1}^{k} p_i$ does not explicitly appear in $f(p_1, p_2, \ldots, p_k)$. For any $i \in [k]$,

$$\frac{\partial f}{\partial p_i} = \frac{\partial}{\partial p_i} P(11123) - \frac{\partial}{\partial q} P(11123) \geq 0,$$

where the inequality follows from Lemma 3.1, and thus Claim 4.2 follows.

We use Claims 4.1 and 4.2 to show that $k = 1$.

- Suppose $k \geq 3$. By Claim 4.2,

$$10p_3^3 - 8p_3^2 + 3p_3 \geq 3p_3 P(11) + 2P(111).$$

Since

$$P(11) \geq p_1^2 + p_2^2 + p_3^2 \geq 0.4549^2 + 2p_3^2,$$
$$P(111) \geq p_1^3 + p_2^3 + p_3^3 \geq 0.4549^3 + 2p_3^3,$$

we have

$$10p_3^3 - 8p_3^2 + 3p_3 \geq 3p_3\left(0.4549^2 + 2p_3^2\right) + 2\left(0.4549^3 + 2p_3^3\right).$$

However, no real number $p_3 \in [0, 1]$ satisfies the above inequality.

- Suppose $k = 2$. Then

$$P(11123) = f(p_1, p_2) = (p_1^3 + p_2^3) - 2(p_1^4 + p_2^4) - (p_1^2 p_2^3 + p_1^3 p_2^2) + (p_1^5 + p_2^5).$$

By the majorization property in Fact 2.3,

$$p_1 + p_2 \leq (3 + 1)/5 < 1.$$

Thus $f(p_1, p_2)$ is not maximized at the boundary $p_1 + p_2 = 1$. Hence

$$\frac{\partial f}{\partial p_1} = \frac{\partial f}{\partial p_2} = 0,$$

i.e.,

$$p_1(5p_1^3 - 8p_1^2 - 3p_1p_2^2 + 3p_1 - 2p_2^3) = 0,$$
$$p_2(5p_2^3 - 8p_2^2 - 3p_2p_1^2 + 3p_2 - 2p_1^3) = 0.$$

Then

$$5p_1^3 - 8p_1^2 - 3p_1p_2^2 + 3p_1 - 2p_2^3 = 5p_2^3 - 8p_2^2 - 3p_2p_1^2 + 3p_2 - 2p_1^3,$$

i.e.,

$$(p_1 - p_2)\left[7(p_1^2 + p_2^2) + 10p_1p_2 - 8(p_1 + p_2) + 3\right] = 0.$$

Since

$$p_1 - p_2 \geq p_1 - (4/5 - p_1) > 2 \cdot 0.4549 - 1 > 0,$$

we have

$$7(p_1^2 + p_2^2) + 10p_1p_2 - 8(p_1 + p_2) + 3 = 0.$$

However,

$$\begin{aligned} & 7(p_1^2 + p_2^2) + 10p_1p_2 - 8(p_1 + p_2) + 3 \\ &= 6\left(p_1 + p_2 - \frac{2}{3}\right)^2 + (p_1 - p_2)^2 + \frac{1}{3} \\ &> 0, \end{aligned}$$

a contradiction.

In conclusion we must have $k = 1$. Then it is easy to show that $p_1 = \frac{3}{5}$. □

4.2.2 Inequalities

We'll break the main proof of Theorem 4.2 into the following steps. In Lemma 4.4, we upper bound $\hat{p}_1$ in terms of $\hat{p}_2$. In Lemma 4.5 we show that $\hat{p}_1$ is close to $\frac{r}{r+u}$, and all other $\hat{p}_i$'s are close to 0. We then prove Theorem 4.2 by showing that Lemmas 4.4 and 4.5 contradict each other if the discrete support size $\hat{k}$ exceeds 1. It follows that k must be one, and the values of $\hat{p}_1$ and $\hat{q}$ can then be calculated.

Lemma 4.4. *For all $r \geq 3$ and $u \geq 2$, $\hat{P}_{1^r23\cdots u+1}$ satisfies, for all $i \in [\hat{k}]$,*

$$u(u-1)\hat{P}_i(1^r23\cdots u-1) \leq r\hat{p}_i^{r-2}\hat{P}_i(12\cdots u) - u\hat{p}_i^{r-1}\hat{P}_i(1^r23\cdots u-1).$$

Furthermore, if $\hat{p}_2 \neq 0$ then

$$u(u-1)\hat{p}_1^r \leq r\hat{p}_2^{r-2}\left[1 + (u-1)\hat{p}_1 - \hat{p}_2\right](1 - \hat{p}_1 - \hat{p}_2).$$

To state Lemma 4.5 we define $L_{r,u}$ and $U_{r,u}$ as in Table 4.1.

Lemma 4.5. *For all $r \geq 3$, if (i) $u = 2$, or (ii) $u > 2$ and $\hat{P}_{1^r23\cdots u} = \left(\frac{r}{r+u-1}\right)$ then $\hat{P}_{1^r23\cdots u+1}$ satisfies*

$$\hat{p}_1 \in \left(U_{r,u}, \frac{r}{r+u}\right],$$

and, for all $i \in [2..\hat{k}]$,

$$\hat{p}_i \in [0, L_{r,u}).$$

Table 4.1: Definitions of $L_{r,u}$ and $U_{r,u}$

$L_{r,u}$	$u=2$	$u=3$	$u\geq 4$
$r=3$	0.3531	0.07869	$\frac{1}{4u}$
$r\geq 4$	$\frac{1}{r}$	$\frac{1}{r+u}$	

$U_{r,u}$	$u=2$	$u=3$	$u\geq 4$
$r=3$	0.4549	0.4199	$\frac{r-1}{r+u}$
$r\geq 4$	$\frac{r-1}{r+u}$		

Lemma 4.4: upper bound of $\hat{p}_1$ using $\hat{p}_2$

We prove the inequalities in Lemma 4.4.

Proof of Lemma 4.4. For simplicity, let $P = \hat{P}_{1^r23\cdots u+1} = (p_1, p_2, \ldots, p_k)$. For convenience, let $p_j = 0$ for any $j > k$. Then for any $i \neq j$,

$$\begin{aligned} P(1^r23\cdots u+1) = {} & P_{i,j}(1^r23\cdots u+1) + u(p_i+p_j)P_{i,j}(1^r23\cdots u) \\ & + u(u-1)p_ip_jP_{i,j}(1^r23\cdots u-1) \\ & + (p_i^r+p_j^r)P_{i,j}(12\cdots u) \\ & + u(p_i^rp_j+p_ip_j^r)P_{i,j}(12\cdots u-1). \end{aligned}$$

Suppose $i \in [k]$ and let $j = k+1$, then $p_i > 0$ and $p_j = 0$. For any $\alpha \in [0,1]$ consider a new distribution P^α, where p_i is replaced by $(1-\alpha)p_i$ and p_j is replaced by $p_j + \alpha p_i = \alpha p_i$. In other words, we split p_i into two probabilities αp_i and $(1-\alpha)p_i$. Then

$$\begin{aligned} & P^\alpha(1^r23\cdots u+1) \\ = {} & P_i(1^r23\cdots u+1) + up_iP_i(1^r23\cdots u) \\ & + u(u-1)(1-\alpha)\alpha p_i^2P_i(1^r23\cdots u-1) \\ & + \left[(1-\alpha)^r+\alpha^r\right]P_i(12\cdots u) \\ & + u\left[(1-\alpha)^r\alpha+\alpha^r(1-\alpha)\right]p_i^{r+1}P_i(12\cdots u-1). \end{aligned}$$

Note that, for $\alpha = 0$, $P^\alpha = P$. Since $\alpha \geq 0$, we have $\frac{\partial}{\partial\alpha}P^\alpha(1^r23\cdots u+1)\big|_{\alpha=0} \leq 0$, *i.e.*

$$u(u-1)p_i^2P_i(1^r23\cdots u-1) \leq rp_i^r\cdot P_i(12\cdots u) - up_i^{r+1}\cdot P_i(12\cdots u-1).$$

Then the first half of Lemma 4.4 follows by removing a factor of p_i^2 on both sides since $p_i > 0$.

To show the second half of Lemma 4.4, if $p_2 > 0$, then the previous inequality implies that

$$u(u-1)P_2(1^r23\cdots u-1) \le rp_2^{r-2}\,P_2(12\cdots u). \tag{4.3}$$

On the other hand, expanding $P_2(1^r23\cdots u-1)$ by p_1, we get

$$\begin{aligned} P_2(1^r23\cdots u-1) = {} & P_{1,2}(1^r23\cdots u-1) \\ & + p_1^r P_{1,2}(12\cdots u-2) + (u-2)p_1P_{1,2}(1^r23\cdots u-2), \end{aligned}$$

where $P_{1,2}(\emptyset) \stackrel{\text{def}}{=} 1$ in case $u = 2$. It follows that

$$P_2(1^r23\cdots u-1) \ge p_1^r \cdot P_{1,2}(12\cdots u-2). \tag{4.4}$$

Furthermore, expanding $P_2(12\cdots u)$ by p_1, we get

$$P_2(12\cdots u) = P_{1,2}(12\cdots u) + up_1P_{1,2}(12\cdots u-1).$$

It is easy to see that

$$\begin{aligned} P_{1,2}(12\cdots u) &\le (1-p_1-p_2)P_{1,2}(1212\cdots u-1), \\ P_{1,2}(12\cdots u-1) &\le (1-p_1-p_2)P_{1,2}(12\cdots u-2). \end{aligned}$$

Thus

$$P_2(12\cdots u) \le (1-p_1-p_2-up_1)(1-p_1-p_2)P_{1,2}(12\cdots u-2). \tag{4.5}$$

Combining Inequalities (4.3), (4.4), and (4.5), we get

$$u(u-1)p_1^rP_2(1^r23\cdots u-2) \le rp_2^{r-2}(1-p_1-p_2-up_1)(1-p_1-p_2)P_{1,2}(12\cdots u-2) \tag{4.6}$$

To cancel $P_{1,2}(12\cdots u-2)$ on both sides of Inequality (4.6), we need to show that $P_{1,2}(12\cdots u-2) > 0$. For $u = 2$ it is clearly true since $P_{1,2}(\emptyset) = 1$. For $u > 2$, suppose that $P_{1,2}(12\cdots u-2) = 0$, then we must have $q = 0$ and $k-2 < u-2$, *i.e.*, $k < u$. We would then have $P(1^r23\cdots u+1) = 0$, contradicting the assumption that $P = \hat{P}_{1^r23\cdots u+1}$. Therefore $P_{1,2}(12\cdots u-2) > 0$ for all $u \ge 2$.

Canceling $P_{1,2}(12\cdots u-2)$ in Inequality (4.6) yields

$$u(u-1)p_1^r \le rp_2^{r-2}\,[1+(u-1)p_1-p_2]\,(1-p_1-p_2). \qquad \square$$

Lemma 4.5: $\hat{p}_i$'s are small or close to $\frac{r}{r+u}$

The proof of Lemma 4.5 consists of the following steps. Consider the ratio

$$\frac{\hat{P}_{1^r23\cdots u}(1^r23\cdots u)}{\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1)}.$$

In Proposition 4.6 we give a lower bound for the ratio, and in Proposition 4.7 we give an upper bound. Using the lower and upper bounds, we prove in Proposition 4.9 that all p_i's in $\hat{P}_{1^r23\cdots u+1}$ are either close to 0 or $\frac{r}{r+u}$, and then Lemma 4.5 follows.

For any $p \in (0,1)$, let

$$F_{r,u}(p) \stackrel{\text{def}}{=} \frac{r-(r+u)p-r(1-p)^{r+u}}{(r-1)u\cdot p(1-p)^{r+u-1}}.$$

Proposition 4.6. *For all $r \geq 3$ and $u \geq 2$, $\hat{P}_{1^r23\cdots u+1}$ satisfies, for all $i \in [\hat{k}]$,*

$$F_{r,u}(\hat{p}_i) \leq \frac{\hat{P}_{1^r23\cdots u}(1^r23\cdots u)}{\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1)}.$$

Proof. For simplicity, let $P = \hat{P}_{1^r23\cdots u+1} = (p_1, p_2, \ldots, p_k)$. For any $i \in [k]$ and $0 \leq \alpha < p_i^{-1}$, consider a new distribution P^α where p_i is scaled by α and all other parts, including q, are scaled by $\frac{1-\alpha p_i}{1-p_i}$. By Expansion (2.3),

$$P(1^r23\cdots u+1) = P_i(1^r23\cdots u+1) + up_iP_i(1^r23\cdots u) + p_i^rP_i(12\cdots u), \quad (4.7)$$

and

$$\begin{aligned} P^\alpha(1^r23\cdots u+1) = &\left(\frac{1-\alpha p_i}{1-p_i}\right)^{r+u} P_i(1^r23\cdots u+1) \\ &+ u(\alpha p_i)\left(\frac{1-\alpha p_i}{1-p_i}\right)^{r+u-1} P_i(1^r23\cdots u) \\ &+ (\alpha p_i)^r\left(\frac{1-\alpha p_i}{1-p_i}\right)^{u} P_i(12\cdots u). \end{aligned}$$

Note that, for $\alpha = 1$, $P^\alpha = P$. Thus, as a function of α, $P^\alpha(1^r23\cdots u+1)$ is maximized at $\alpha = 1$. Note that $p_i^{-1} > 1$. Then $\alpha = 1$ is not at the boundary. Thus

$$\frac{\partial}{\partial\alpha}P^\alpha(1^r23\cdots u+1)\bigg|_{\alpha=1} = 0,$$

i.e.,

$$0 = -(r+u)P_i(1^r23\cdots u+1) + u\left[1-(r+u)p_i\right]P_i(1^r23\cdots u) + p_i^{r-1}\left[r-(r+u)p_i\right]P_i(12\cdots u). \quad (4.8)$$

Eliminating $P_i(12\cdots u)$ from Equations (4.7) and (4.8) yields

$$\left[r-(r+u)p_i\right]P(1^r23\cdots u+1) = rP_i(1^r23\cdots u+1) + (r-1)up_iP_i(1^r23\cdots u).$$

Let P_i^{norm} be the distribution obtain from P by removing p_i and normalizing the remaining probabilities. Then

$$\begin{aligned} P_i^{\text{norm}}(1^r23\cdots u+1) &\le (1-p_i)^{r+u}P_i^{\text{norm}}(1^r23\cdots u+1) \\ &\le (1-p_i)^{r+u}\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1), \end{aligned}$$

and similarly

$$P_i^{\text{norm}}(1^r23\cdots u) \le (1-p_i)^{r+u}\hat{P}_{1^r23\cdots u}(1^r23\cdots u).$$

Thus

$$\begin{aligned} \left[r-(r+u)p_i\right]P(1^r23\cdots u+1) \le{}& r(1-p_i)^{r+u}\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1) \\ &+ (r-1)up_i(1-p_i)^{r+u}\hat{P}_{1^r23\cdots u}(1^r23\cdots u), \end{aligned}$$

which can be rewritten as

$$F_{r,u}(p_i) \le \frac{\hat{P}_{1^r23\cdots u}(1^r23\cdots u)}{\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1)}. \qquad \square$$

Let

$$A_{r,u} \stackrel{\text{def}}{=} \left(\frac{r+u}{r}\right)^r\left(\frac{r+u}{u}\right)^u \cdot \begin{cases} \frac{1}{8}, & \text{if } r=3 \text{ and } u=2, \\ \frac{2}{5r}, & \text{if } r\ge 4 \text{ and } u=2, \\ \left(\frac{r}{r+u-1}\right)^r\left(\frac{u-1}{r+u-1}\right)^{u-1}, & \text{if } r\ge 3 \text{ and } u\ge 3. \end{cases}$$

Proposition 4.7. *For all $r \ge 3$, if (i) $u=2$, or (ii) $u>2$ and $\hat{P}_{1^r23\cdots u} = \left(\frac{r}{r+u-1}\right)$, then $\hat{P}_{1^r23\cdots u+1}$ satisfies, for all $i \in [\hat{k}]$,*

$$\frac{\hat{P}_{1^r23\cdots u}(1^r23\cdots u)}{\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1)} \le A_{r,u}.$$

Proof. Note that in the ratio $\frac{\hat{P}_{1^r23\cdots u}(1^r23\cdots u)}{\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1)}$ the denominator can be bounded by

$$P_{1^r23\cdots u+1}(1^r23\cdots u+1) \geq \left(\frac{r+u}{r}\right)^r \left(\frac{r+u}{u}\right)^u.$$

Furthermore, if $u > 2$, by assumption

$$P_{1^r23\cdots u}(1^r23\cdots u) = \left(\frac{r}{r+u-1}\right)^r \left(\frac{u-1}{r+u-1}\right)^{u-1}.$$

Examining the definition of $A_{r,u}$, we only need to show that

$$\hat{P}_{1^r2}(1^r2) \leq \begin{cases} \frac{1}{8}, & \text{if } r = 3, \\ \frac{2}{5r}, & \text{if } r \geq 4. \end{cases}$$

For $u = 2$, the pattern $1^r23\cdots u = 1^r2$ becomes binary. As described in Chapter 2, the PML distribution of any non-trivial binary pattern has support size 2, and the probabilities can be found by solving an uni-variate equation. As for 1^r2, the PML distribution is $\hat{P}_{1^r2} = (p, 1-p)$, where p satisfies

$$\left[rp^{r-1}(1-p) - p^r\right] + \left[(1-p)^r - r(1-p)^{r-1}\right] = 0.$$

For $r = 3$, it is easy to verify that $\hat{P}_{1112} = \left(\frac{1}{2}, \frac{1}{2}\right)$ and $\hat{P}_{1112}(1112) = 1/8$. For $r > 3$, we can show that

Claim 4.3. *For all $r > 3$, $\hat{P}_{1^r2}(1^r2) \leq \frac{2}{5r}$.*

see Appendix A.1 for the complete proof of Claim 4.3. Combining all cases we have that

$$\frac{\hat{P}_{1^r23\cdots u}(1^r23\cdots u)}{\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1)} \leq A_{r,u}. \qquad \square$$

Combining the lower bound in Propositions 4.6 and the upper bound in 4.7, we get

Proposition 4.8. *For all $r \geq 3$, if (i) $u = 2$, or (ii) $u > 2$ and $\hat{P}_{1^r23\cdots u} = \left(\frac{r}{r+u-1}\right)$, then $\hat{P}_{1^r23\cdots u+1}$ satisfies, for all $i \in [\hat{k}]$,*

$$F_{r,u}(\hat{p}_i) \leq A_{r,u}.$$

On the other hand, we show the following property of $F_{r,u}(p)$:

Claim 4.4. *Given any $r \geq 3$ and $u \geq 2$, $F_{r,u}(p) > A_{r,u}$ for all $p \in (U_{r,u}, L_{r,u})$.*

See Appendix A.1 for the complete proof of Claim 4.4. Combined with Proposition 4.8 it implies that no $\hat{p}_i$ in $\hat{P}_{1^r23\cdots u+1}$ lies in the interval $[L_{r,u}, U_{r,u}]$. Note that the majorization property in Fact 2.3 implies that any $\hat{p}_i$ is at most $\frac{r}{r+u}$. Therefore we have the following:

Proposition 4.9. *For all $r \geq 3$, if (i) $u = 2$, or (ii) $u > 2$ and $\hat{P}_{1^r23\cdots u} = \left(\frac{r}{r+u-1}\right)$, then $\hat{P}_{1^r23\cdots u+1}$ satisfies, for all $i \geq 1$,*

$$\hat{p}_i \in (0, L_{r,u}) \bigcup \left(U_{r,u}, \frac{r}{r+u}\right].$$

Finally we can prove Lemma 4.5.

Proof of Lemma 4.5. For simplicity, let $P = \hat{P}_{1^r23\cdots u+1} = (p_1, p_2, \ldots, p_k)$. By Proposition 4.9, for all $i \in [k]$, $p_i \in (0, L_{r,u})$ or $p_i \in \left(U_{r,u}, \frac{r}{r+u}\right]$. It is then sufficient to prove that $p_1 \geq L_{r,u}$ and, for all $i \in [2..k]$, $p_i \leq U_{r,u}$.

To show that $p_1 \geq L_{r,u}$, note that

$$P(1^r23\cdots u+1) \geq \left(\frac{r}{r+u}\right)^r \left(\frac{u}{r+u}\right)^u,$$

and

$$P(1^r23\cdots u+1) \leq p_1^{r-2} P(1123\cdots u+1) \leq p_1^{r-2} \cdot \hat{P}_{1123\cdots u+1}(1123\cdots u+1).$$

Then

$$p_1^{r-2} \geq \frac{\hat{P}_{1123\cdots u+1}(1123\cdots u+1)}{\left(\frac{r}{r+u}\right)^r \left(\frac{u}{r+u}\right)^u}. \tag{4.9}$$

Note that $1123\cdots u+1$ is 1-uniform. As mentioned in 2, the probability of a 1-uniform pattern can be maximized at a uniform distribution, and the support size k_1 can be found as $\arg\min_{k_1 \geq m} k_1 \geq m\frac{k_1+1}{k_1-(u+1)+1} \cdot \left(\frac{k_1}{k_1+1}\right)^{u+2} \leq 1$.

For $r = 3$ and $u = 2$, it is easy to find that $\hat{P}_{1123}$ is uniform with support size $k_1 = 5$. Thus, for $r = 3$ and $u = 2$, Inequality 4.9 implies that

$$p_1 \geq \frac{\hat{P}_{1123}(1123)}{\left(\frac{3}{5}\right)^3 \left(\frac{2}{5}\right)^2} \geq \frac{9}{25} = 0.36 > L_{3,2}.$$

For $r > 3$ or $u > 2$, it can be shown that

Claim 4.5. *For all $m \geq 2$,*

$$\hat{P}_{1123\cdots m}(1123\cdots m) \leq \frac{2/e}{m(m-1)}.$$

Furthermore,

Claim 4.6. *If $r \geq 3$ and $u \geq 2$, but $r \neq 3$ or $u \neq 2$, then*

$$\left(\frac{r}{r+u}\right)^r \left(\frac{u}{r+u}\right)^u \cdot \frac{u(u+1)}{2} > L_{r,u}^{r-2}.$$

See Appendix A.1 for the complete proofs of Claims 4.5 and 4.6. Combining Inequality (4.9) with these two claims, we get $p_1 > L_{r,u}$.

In either case, we have $p_1 > L_{r,u}$. Hence

$$p_1 \in \left(U_{r,u}, \frac{r}{r+u}\right].$$

To show the second half of Lemma 4.5, by the majorization property in Fact 2.3, for all $i \in [2..k]$, $p_1 + p_i \leq p_1 + p_2 \leq \frac{r+1}{r+u}$. Therefore

$$p_i \leq \frac{r+1}{2(r+u)} \leq U_{r,u},$$

where the second inequality can be directly verified by calculation for all $r \geq 3$ and $u \geq 2$. It follows that, for all $i \in [2..k]$, $p_i \in (0, L_{r,u})$. □

4.2.3 Proof for Skewed Patterns

We use Theorem 4.3 and Lemmas 4.4 and 4.5 to complete the proof of Theorem 4.2.

Proof of Theorem 4.2. For simplicity, let $P = \hat{P}_{1^r23\cdots u+1} = (p_1, p_2, \ldots, p_k)$. We use induction on $u \geq 2$ to show that $k = 1$. Then it is easy to show that $p_1 = \frac{r}{r+u}$.
Basis ($\boldsymbol{u = 2}$) As show in Theorem 4.3, for $u = 2$ and $r = 3$, $\hat{P}_{11123} = \left(\frac{3}{5}\right)$. For $u = 2$ and $r > 3$,we show that $k = 1$ then it is easy to show that $p_1 = \frac{r}{r+u}$.

Suppose $k \geq 2$, then $p_2 > 0$. By Lemmas 4.4 and 4.5 we have

$$2p_1^r \leq rp_2^{r-2}\,(1 + p_1 - p_2)\,(1 - p_1 - p_2),$$
$$p_1 \in \left(U_{r,2}, \frac{r}{r+2}\right].$$

Then

$$2U_{r,2}^r \leq rp_2^{r-2}\left[\frac{2(r+1)}{r+2} - p_2\right](1 - U_{r,2} - p_2)\,. \tag{4.10}$$

Lemma 4.5 also says that $p_2 < L_{r,u}$. Furthermore, we can show that the right-hand side of Inequality 4.10 increases in p_2:

Claim 4.7. *If $r \geq 3$ and $u \geq 2$, but $r \neq 3$ or $u \neq 2$, then*

$$p^{r-2}\left[\frac{(r+1)u}{r+u} - p\right](1 - U_{r,u} - p)$$

increases for $p \in (0, L_{r,u}]$.

See Appendix A.1 for the complete proof of Claim 4.7. Substituting $L_{r,u}$ for p_2 in Inequality 4.10, we get

$$2U_{r,2}^r \leq rL_{r,2}^{r-2}\left[\frac{2(r+1)}{r+2} - L_{r,2}\right](1 - U_{r,2} - L_{r,2})\,. \tag{4.11}$$

However, we can verify case by case that one can show that the left-hand side of Inequality (4.11) is always larger than the right-hand side:

Claim 4.8. *If $r \geq 3$ and $u \geq 2$, but $r \neq 3$ or $u \neq 2$, then*

$$u(u-1)U_{r,u}^r > rL_{r,u}^{r-2}\left[\frac{(r+1)u}{r+u} - L_{r,u}\right](1 - U_{r,u} - L_{r,u}).$$

See Appendix A.1 for the complete proof of Claim 4.8. Thus Inequality (4.11) and Claim 4.8 contradict each other. Thus our assumption that $k \geq 2$ is false and we must have $k = 1$, completing the proof of the basis.

Induction Step We show that for all $u > 2$ and $r \geq 3$, if $\hat{P}_{1^r 23\cdots u} = \left(\frac{r}{r+u-1}\right)$, then $\hat{P}_{1^r 23\cdots u+1} = \left(\frac{r}{r+u}\right)$.

The proof resembles that of the basis. We prove by contradiction that $k = 1$. Suppose $k \geq 2$ then $p_2 > 0$. By Lemmas 4.4 and 4.5 we have

$$u(u-1)p_1^r \leq rp_2^{r-2}\,[1 + (u-1)p_1 - p_2]\,(1 - p_1 - p_2),$$
$$p_1 \in \left(U_{r,u}, \frac{r}{r+u}\right].$$

Then

$$u(u-1)U_{r,u}^r \le rp_2^{r-2}\left[\frac{(r+1)u}{r+u} - p_2\right](1 - U_{r,u} - p_2)\,.$$

Lemma 4.5 also says that $p_2 < L_{r,u}$. Using Claim 4.7, we get

$$u(u-1)U_{r,u}^r \le rL_{r,u}^{r-2}\left[\frac{(r+1)u}{r+u} - L_{r,u}\right](1 - U_{r,u} - L_{r,u})\,,$$

which contradicts Claim 4.8. It follows that our assumption $k \ge 2$ does not hold; we must have $k = 1$.

Combining both the basis and induction step, we have that ,for all $r \ge 3$ and $u \ge 2$, $\hat{P}_{1^r23\cdots u+1}$ has discrete support size $k = 1$. It follows that $P(1^r23\cdots u+1)$ is $p_1^r(1-p_1)^u$, which is maximized at $p_1 = \hat{p}_1 = \frac{r}{r+u}$. □

4.3 Quasi-uniform Patterns

Recall that $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_m$ denote the multiplicities of a canonical pattern. In a *uniform* pattern $\mu_t = \mu_{t'}$ for all $t, t' \in [m]$. In a *1-uniform* pattern $|\mu_t - \mu_{t'}| \le 1$ for all $t, t' \in [m]$. As stated in Facts 2.10 and 2.11, the PML distributions of uniform and 1-uniform patterns are essentially uniform. We generalize the results by relaxing the bound constraint on the multiplicities. Let $\binom{S}{i}$ be the set of i-element subsets of a set S. For any $\{t, t'\} \in \binom{[m]}{2}$, let

$$d_{t,t'} \stackrel{\text{def}}{=} (\mu_t - \mu_{t'})^2 - (\mu_t + \mu_{t'} - 2).$$

In a *quasi-uniform* pattern $d_{t,t'} \le 0$ for all $\{t, t'\} \in \binom{[m]}{2}$. We show that

Theorem 4.10. *The PML of any non-trivial quasi-uniform $\bar{\psi}$ can be achieved at a uniform distribution.*

We first prove an algebraic inequality.

Lemma 4.11. *For any real numbers $p \ne q$ and integer $n \ge 0$,*

$$\frac{p^n - q^n}{p - q} \le \frac{n}{2}\left(p^{n-1} + q^{n-1}\right),$$

where the equality holds if and only if $n \in \{0, 1\}$.

Proof. We first rewrite the left-hand side as

$$\frac{p^n - q^n}{p - q} = \sum_{i=0}^{n-1} p^i q^{n-1-i}.$$

Note that for any $i = 0, 1, \ldots, n-1$,

$$\left(p^i q^{n-1-i} + p^{n-1-i} q^i\right) - \left(p^{n-1} + q^{n-1}\right) = -\left(p^i - q^i\right)\left(p^{n-1-i} - q^{n-1-i}\right) \leq 0.$$

Then

$$\frac{p^n - q^n}{p - q} \leq \frac{1}{2} \sum_{i=0}^{n-1} \left(p^{n-1} + q^{n-1}\right) = \frac{n}{2}\left(p^{n-1} + q^{n-1}\right). \qquad \square$$

Recall that, given pattern $\bar{\psi}$ and $K \geq m$, the bounded PML distribution is

$$\hat{P}_{\bar{\psi}}^{(K)} \stackrel{\text{def}}{=} \underset{P \in \mathcal{P}_d^{\text{sorted}}: |P| \leq K}{\arg\max} P(\bar{\psi}),$$

and

$$\lim_{K \to \infty} \hat{P}^{(K)}(\bar{\psi}) = \hat{P}(\bar{\psi}).$$

If for every $K \geq m$, $\hat{P}_{\bar{\psi}}^{(K)}$ is uniform, similar to finding the support size of the PML distribution for uniform or 1-uniform patterns, $\lim_{K\to\infty} \hat{P}^{(K)}(\bar{\psi})$ can be achieved at the uniform distribution with support

$$\hat{k} = \underset{k \geq m}{\arg\min} \frac{k+1}{k-m+1} \cdot \left(\frac{k}{k+1}\right)^n \leq 1.$$

Thus to prove Theorem 4.10 it's sufficient to show that, for any $K \geq m$, the bounded PML distribution $\hat{P}_{\bar{\psi}}^{(K)}$ is uniform. Without loss of generality, we assume there is a fixed K, and $\hat{P}_{\bar{\psi}}^{(K)} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k) \in \mathcal{P}_d^{\text{sorted}}$.

Proof of Theorem 4.10. For simplicity, let $P = (p_1, p_2, \ldots, p_k) = \hat{P}_{\bar{\psi}}^{(K)}$ be the bounded PML distribution. Furthermore, without loss of generality we assume that there is no other bounded PML distribution with smaller support size.

Suppose P is not uniform *i.e.*, $p_1 > p_k$. By Expansion (2.6),

$$P(\bar{\psi}) = P_{1,k}(\bar{\psi}) + \sum_t (p_1^{\mu_t} + p_k^{\mu_t}) P_{1,k}(\bar{\psi}_t) + \sum_{(t,t') \in [m]^{\underline{2}}} p_1^{\mu_t} p_k^{\mu_{t'}} P_{1,k}(\bar{\psi}_{t,t'}),$$

where $[m]^{\underline{2}}$ is the set of ordered pairs (t, t') with $t, t' \in [m]$ such that $t \neq t'$. Let P' be the distribution with p_1 and p_k merged, *i.e.*, replacing p_1 and p_k by $p'_1 = p_1 + p_k$ and $p'_k = 0$. Then $P'(\bar{\psi})$ can be expanded as

$$P'(\bar{\psi}) = P_{1,k}(\bar{\psi}) + \sum_t (p_1 + p_k)^{\mu_t} P_{1,k}(\bar{\psi}_t).$$

Since we assumed that no other bounded PML distribution has smaller support size, $P'(\bar{\psi})$ is strictly less than $P(\bar{\psi})$, and hence combining the expansions for $P(\bar{\psi})$ and $P'(\bar{\psi})$ we get

$$\sum_{t:\mu_t \geq 2} \left[(p_1 + p_k)^{\mu_t} - (p_1^{\mu_t} + p_k^{\mu_t})\right] P_{1,k}(\bar{\psi}_t) < \sum_{(t,t') \in [m]^{\underline{2}}} p_1^{\mu_t} p_k^{\mu_{t'}} P_{1,k}(\bar{\psi}_{t,t'}).$$

Note that, for any $\mu_t \geq 2$,

$$(p_1 + p_k)^{\mu_t} - (p_1^{\mu_t} + p_k^{\mu_t}) \geq \binom{\mu_t}{2} p_1 p_k (p_1^{\mu_t - 2} + p_k^{\mu_t - 2}).$$

Then

$$\sum_{t:\mu_t \geq 2} \binom{\mu_t}{2} (p_1^{\mu_t - 2} + p_k^{\mu_t - 2}) P_{1,k}(\bar{\psi}_t) < \sum_{(t,t') \in [m]^{\underline{2}}} p_1^{\mu_t - 1} p_k^{\mu_{t'} - 1} P_{1,k}(\bar{\psi}_{t,t'}). \tag{4.12}$$

On the other hand, let P'' be the distribution with p_1 and p_k replaced by $p''_1 = \frac{p_1 + p_k}{2}$ and $p''_k = \frac{p_1 + p_k}{2}$. Then $P''(\bar{\psi})$ can be expanded as

$$P''(\bar{\psi}) = P_{1,k}(\bar{\psi}) + \sum_t 2 \left(\frac{p_1 + p_k}{2}\right)^{\mu_t} P_{1,k}(\bar{\psi}_t) + \sum_{(t,t')} \left(\frac{p_1 + p_k}{2}\right)^{\mu_t + \mu_{t'}} P_{1,k}(\bar{\psi}_{t,t'}).$$

Combining the expansions for $P(\bar{\psi})$ and $P''(\bar{\psi})$, we get

$$\sum_{(t,t') \in [m]^{\underline{2}}} \left[\left(\frac{p_1 + p_k}{2}\right)^{\mu_t + \mu_{t'}} - p_1^{\mu_t} p_k^{\mu_{t'}}\right] P_{1,k}(\bar{\psi}_{t,t'}) \\ \leq \sum_{t:\mu_t \geq 2} \left[(p_1^{\mu_t} + p_k^{\mu_t}) - 2\left(\frac{p_1 + p_k}{2}\right)^{\mu_t}\right] P_{1,k}(\bar{\psi}_t). \tag{4.13}$$

Note that

$$p_1^{\mu_t} - \left(\frac{p_1 + p_k}{2}\right)^{\mu_t} = \frac{p_1 - p_k}{2} \sum_{\ell=0}^{\mu_t - 1} p_1^{\ell} \left(\frac{p_1 + p_k}{2}\right)^{\mu_t - 1 - \ell}$$

$$p_k^{\mu_t} - \left(\frac{p_1 + p_k}{2}\right)^{\mu_t} = -\frac{p_1 - p_k}{2} \sum_{\ell=0}^{\mu_t - 1} p_k^{\ell} \left(\frac{p_1 + p_k}{2}\right)^{\mu_t - 1 - \ell}$$

It follows that

$$(p_1^{\mu_t} + p_k^{\mu_t}) - 2\left(\frac{p_1+p_k}{2}\right)^{\mu_t} = \frac{p_1-p_k}{2}\sum_{\ell=1}^{\mu_t-1}\left(p_1^{\ell} - p_k^{\ell}\right)\left(\frac{p_1+p_k}{2}\right)^{\mu_t-1-\ell}.$$

Since $p_1 \neq p_k$, by Lemma 4.11, for any $\ell = 1, 2, \ldots, \mu_t - 1$,

$$\frac{p_1^{\ell} - p_k^{\ell}}{p_1 - p_k} \leq \frac{\ell}{2}\left(p_1^{\ell} + p_k^{\ell}\right).$$

Furthermore, since $x^{\mu_t-1-\ell}$ is a convex function,

$$\left(\frac{p_1+p_k}{2}\right)^{\mu_t-1-\ell} \leq \frac{1}{2}\left(p_1^{\mu_t-1-\ell} + p_k^{\mu_t-1-\ell}\right).$$

Thus

$$(p_1^{\mu_t} + p_k^{\mu_t}) - 2\left(\frac{p_1+p_k}{2}\right)^{\mu_t} \leq \frac{(p_1-p_k)^2}{2}\sum_{\ell=1}^{\mu_t-1}\frac{\ell}{2}(p_1^{\ell-1} + p_k^{\ell-1})\cdot\frac{1}{2}(p_1^{\mu_t-1-\ell} + p_k^{\mu_t-1-\ell}),$$

where

$$\begin{aligned}(p_1^{\ell-1} + p_k^{\ell-1})(p_1^{\mu_t-1-\ell} + p_k^{\mu_t-1-\ell}) &= \left(p_1^{\mu_t-2} + p_k^{\mu_2-2}\right) + \left(p_1^{\ell-1}p_k^{\mu_t-1-\ell} + p_1^{mu_t-1-\ell}p_k^{\ell-1}\right)\\ &\leq 2\left(p_1^{\mu_t-2} + p_k^{\mu_2-2}\right).\end{aligned}$$

Then

$$\begin{aligned}(p_1^{\mu_t} + p_k^{\mu_t}) - 2\left(\frac{p_1+p_k}{2}\right)^{\mu_t} &\leq \frac{(p_1-p_k)^2}{2}\sum_{\ell=1}^{\mu_t-1}\frac{\ell}{2}\left(p_1^{\mu_t-2} + p_k^{\mu_2-2}\right)\\ &= \frac{(p_1-p_k)^2}{4}\left(p_1^{\mu_t-2} + p_k^{\mu_2-2}\right)\cdot\frac{\mu_t(\mu_t-1)}{2},\end{aligned}$$

which implies, combined with Inequality (4.13), that

$$\begin{aligned}&\sum_{(t,t')\in[m]^{\underline{2}}}\left[\left(\frac{p_1+p_k}{2}\right)^{\mu_t+\mu_{t'}} - p_1^{\mu_t}p_k^{\mu_{t'}}\right]P_{1,k}(\bar{\psi}_{t,t'})\\ &\qquad\leq \frac{(p_1-p_k)^2}{4}\sum_{t:\mu_t\geq 2}\binom{\mu_t}{2}(p_1^{\mu_t-2} + p_k^{\mu_t-2})P_{1,k}(\bar{\psi}_t). \quad (4.14)\end{aligned}$$

Eliminating $\sum_{t:\mu_t\geq 2}\binom{\mu_t}{2}(p_1^{\mu_t-2} + p_k^{\mu_t-2})P_{1,k}(\bar{\psi}_t)$ from Inequalities (4.12) and (4.14),

$$\begin{aligned}&\sum_{(t,t')\in[m]^{\underline{2}}}\left[\left(\frac{p_1+p_k}{2}\right)^{\mu_t+\mu_{t'}} - p_1^{\mu_t}p_k^{\mu_{t'}}\right]P_{1,k}(\bar{\psi}_{t,t'})\\ &\qquad< \frac{(p_1-p_k)^2}{4}\sum_{(t,t')\in[m]^{\underline{2}}}p_1^{\mu_t}p_k^{\mu_{t'}}P_{1,k}(\bar{\psi}_{t,t'}),\end{aligned}$$

which can be rewritten as

$$\sum_{\{t,t'\}\in\binom{[m]}{2}} \left[2\left(\frac{p_1+p_k}{2}\right)^{\mu_t+\mu_{t'}-2} - \left(p_1^{\mu_t-1}p_k^{\mu_{t'}-1} + p_1^{\mu_{t'}-1}p_k^{\mu_t-1}\right)\right] P_{1,k}(\bar{\psi}_{t,t'}) < 0.$$

Let $r = \frac{p_1}{p_k} > 1$. Then

$$\sum_{\{t,t'\}\in\binom{[m]}{2}} \left[\frac{2\left(\frac{r+1}{2}\right)^{\mu_t+\mu_{t'}-2}}{r^{\mu_t-1}+r^{\mu_{t'}-1}} - 1\right] \left(p_1^{\mu_t}p_k^{\mu_{t'}} + p_1^{\mu_{t'}}p_k^{\mu_t}\right) P_{1,k}(\bar{\psi}_{t,t'}) \leq 0. \tag{4.15}$$

However, by Theorem 4.1,

$$\left(\frac{r}{r+1}\right)^{\mu_t-1}\left(\frac{1}{r+1}\right)^{\mu_{t'}-1} + \left(\frac{1}{r+1}\right)^{\mu_t-1}\left(\frac{r}{r+1}\right)^{\mu_{t'}-1} \leq 2\left(\frac{1}{2}\right)^{(\mu_t-1)+(\mu_{t'}-1)},$$

i.e.,

$$r^{\mu_t-1} + r^{\mu_{t'}-1} \leq 2\left(\frac{r+1}{2}\right)^{\mu_t+\mu_{t'}-2},$$

which contradicts Inequality (4.15). Therefore our assumption that $P = \hat{P}_{\bar{\psi}}^{(K)}$ is non-uniform is false and the conclusion follows. □

4.4 Almost-uniform Patterns

A pattern $\bar{\psi}$ is *almost-uniform* if for all $t, t' \in [m]$,

$$(\mu_1-1)^{\mu_1-\mu_m}\sum_{d_{t,t'}>0} d_{t,t'} + (\mu_m-1)^{\mu_1-\mu_m}\sum_{d_{t,t'}<0} d_{t,t'} \leq 0,$$

where we consider $0^0 = 1$ in the case $\mu_1 = \mu_m = 1$.

Note that a quasi-uniform pattern is also almost-uniform. Particularly, if $\mu_m = 1$, then $\bar{\psi}$ is almost-uniform only if $\sum_{d_{t,t'}>0} d_{t,t'} = 0$, *i.e.*, the almost-uniform pattern degenerates to a quasi-uniform pattern.

In general, if $\mu_m > 1$, the condition for a pattern being almost-uniform can be rewritten as

$$\frac{\sum_{d_{t,t'}<0}|d_{t,t'}|}{\sum_{d_{t,t'}>0} d_{t,t'}} \geq \left(\frac{\mu_1-1}{\mu_m-1}\right)^{\mu_1-\mu_m}.$$

We show in Theorem 4.12 that the PML distribution of an almost-uniform pattern is essentially uniform. We also show similar results in Theorem 4.13 for certain patterns with largest multiplicity $\mu_1 = 3$ but not necessarily almost-uniform.

Theorem 4.12. *The PML of any non-trivial almost-uniform $\bar{\psi}$ can be achieved at a uniform distribution.*

Theorem 4.13. *For any pattern $\bar{\psi}$ such that $\mu_1 = 3$ and*

$$4\binom{\varphi_2}{2} \geq \varphi_1\varphi_3,$$

the PML of $\bar{\psi}$ can be achieved at a uniform distribution.

Similar to the proof of Theorem 4.10, it's sufficient to consider only bounded PML distribution. Without loss of generality, we assume there is a fixed K, and $\hat{P}^{(K)}_{\bar{\psi}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k) \in \mathcal{P}^{\text{sorted}}_d$.

We first show an inequality derived from merging two probabilities in the PML distribution. We'll use this inequality in the proofs of both Theorems 4.12 and 4.13. The structure of the proofs is illustrated in Figure 4.2.

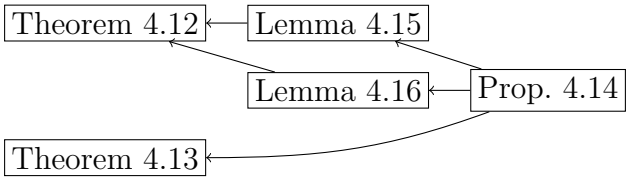


Figure 4.2: Roadmap to the proof for Almost-uniform Patterns

Proposition 4.14. *Let $\hat{p}_i > \hat{p}_j > 0$ be probabilities from a (bounded) PML distribution $\hat{P}_{\bar{\psi}}$ of pattern $\bar{\psi}$. Let $\hat{P}'$ be the distribution obtained from $\hat{P}_{\bar{\psi}}$ by replacing $\hat{p}_i$ and $\hat{p}_j$ with $\hat{p}'_i \stackrel{\text{def}}{=} \hat{p}_i + \hat{p}_j$ and $\hat{p}'_j = 0$ respectively. Then*

$$\hat{p}_i\hat{p}_j \sum_{\{t,t'\}\in\binom{[m]}{2}} c_{t,t'}\hat{P}_{i,j}(\bar{\psi}_{t,t'}) \geq \hat{P}_{\bar{\psi}}(\bar{\psi}) - \hat{P}'(\bar{\psi}) \geq 0.$$

where

$$c_{t,t'} = (\mu_t - 1)\frac{\hat{p}_i^{\mu_t-1}\hat{p}_j^{\mu_{t'}} - \hat{p}_i^{\mu_{t'}}\hat{p}_j^{\mu_t-1}}{\hat{p}_i - \hat{p}_j} + (\mu_{t'} - 1)\frac{\hat{p}_i^{\mu_{t'}-1}\hat{p}_j^{\mu_t} - \hat{p}_i^{\mu_t}\hat{p}_j^{\mu_{t'}-1}}{\hat{p}_i - \hat{p}_j}.$$

Observation: The second inequality is strict if there is no (bounded) PML distribution with smaller support size.

Proof. For simplicity, let $P = (p_1, p_2, \ldots, p_k)$, either a PML distribution or a bounded PML distribution. By Expansion (2.6), setting $I = \{i, j\}$, the pattern probability $P(\bar{\psi})$ can be written as

$$\begin{aligned} P(\bar{\psi}) &= P_{i,j}(\bar{\psi}) + \sum_{t=1}^{m}(p_i^{\mu_t} + p_j^{\mu_t})P_{i,j}(\bar{\psi}_t) \\ &\quad + \sum_{\{t,t'\}\in\binom{[m]}{2}} (p_i^{\mu_t}p_j^{\mu_{t'}} + p_i^{\mu_{t'}}p_j^{\mu_t})P_{i,j}(\bar{\psi}_{t,t'}), \qquad (4.16) \\ P'(\bar{\psi}) &= P_{i,j}(\bar{\psi}) + \sum_{t=1}^{m}(p_i + p_j)^{\mu_t}P_{i,j}(\bar{\psi}_t). \end{aligned}$$

Taking the difference, we get

$$\begin{aligned} P(\bar{\psi}) - P'(\bar{\psi}) = &\sum_{\{t,t'\}\in\binom{[m]}{2}} (p_i^{\mu_t}p_j^{\mu_{t'}} + p_i^{\mu_{t'}}p_j^{\mu_t})P_{i,j}(\bar{\psi}_{t,t'}) \\ &- \sum_{t=1}^{m}\left[(p_i + p_j)^{\mu_t} - (p_i^{\mu_t} + p_j^{\mu_t})\right]P_{i,j}(\bar{\psi}_t). \end{aligned}$$

Note that, for all integers $\mu \geq 1$,

$$(p_i + p_j)^{\mu} - (p_i^{\mu} + p_j^{\mu}) = \sum_{t=1}^{\mu-1}\binom{\mu}{t}p_i^t p_j^{\mu-t} \geq \mu\sum_{t=1}^{\mu-1}p_i^t p_j^{\mu-t} = \mu(p_ip_j)\frac{p_i^{\mu-1} - p_j^{\mu-1}}{p_i - p_j},$$

where the second equality holds if and only if $\mu \leq 3$. Then

$$P(\bar{\psi}) - P'(\bar{\psi}) \leq p_i p_j \sum_{\{t,t'\} \in \binom{[m]}{2}} (p_i^{\mu_t - 1} p_j^{\mu_{t'} - 1} + p_i^{\mu_{t'} - 1} p_j^{\mu_t - 1})(p_i - p_j) P_{i,j}(\bar{\psi}_{t,t'}) \\ - p_i p_j \sum_{t=1}^{m} \mu_t (p_i^{\mu_t - 1} - p_j^{\mu_t - 1}) P_{i,j}(\bar{\psi}_t), \quad (4.17)$$

where the equality holds only if $\mu_t \leq 3$ for all $t \in [m]$. On the other hand, by Expansion (4.16),

$$\frac{\partial P(\bar{\psi})}{\partial p_i} = \sum_{t=1}^{m} \mu_t p_i^{\mu_t - 1} P_{i,j}(\bar{\psi}_t) + \sum_{\{t,t'\} \in \binom{[m]}{2}} (\mu_t p_i^{\mu_t - 1} p_j^{\mu_{t'}} + \mu_{t'} p_i^{\mu_{t'} - 1} p_j^{\mu_t}) P_{i,j}(\bar{\psi}_{t,t'}),$$

$$\frac{\partial P(\bar{\psi})}{\partial p_j} = \sum_{t=1}^{m} \mu_{t'} p_j^{\mu_{t'}} P_{i,j}(\bar{\psi}_t) + \sum_{\{t,t'\} \in \binom{[m]}{2}} (\mu_{t'} p_i^{\mu_t} p_j^{\mu_{t'} - 1} + \mu_t p_i^{\mu_{t'}} p_j^{\mu_t - 1}) P_{i,j}(\bar{\psi}_{t,t'}).$$

By Lemma 3.1, $\frac{\partial P(\bar{\psi})}{\partial p_i} = \frac{\partial P(\bar{\psi})}{\partial p_j}$. Then

$$\sum_{t=1}^{m} \mu_t (p_i^{\mu_t - 1} - p_j^{\mu_t - 1}) P_{i,j}(\bar{\psi}_t) + \sum_{\{t,t'\}: t \neq t'} \Big[\mu_t (p_i^{\mu_t - 1} p_j^{\mu_{t'}} - p_i^{\mu_{t'}} p_j^{\mu_t - 1}) \\ + \mu_{t'} (p_i^{\mu_{t'} - 1} p_j^{\mu_t} - p_i^{\mu_t} p_j^{\mu_{t'} - 1}) \Big] P_{i,j}(\bar{\psi}_{t,t'}) = 0. \quad (4.18)$$

The conclusion follows by eliminating $\sum_{t=1}^{m} \mu_t (p_i^{\mu_t - 1} - p_j^{\mu_t - 1}) P_{i,j}(\bar{\psi}_t)$ from Inequality (4.17) and Equation (4.18). □

4.4.1 Theorem 4.12: Almost-uniform Patterns

The proof of Theorem 4.12 consists of two steps. In Lemma 4.15 we show that both the PML distribution and bounded PML distribution of any non-trivial pattern satisfies $\frac{\hat{p}_1}{\hat{p}_k} \leq \frac{\mu_1 - 1}{\mu_m - 1}$. In Lemma 4.16 we show that the bounded PML distribution is uniform if $\left(\frac{\hat{p}_1}{\hat{p}_k}\right)^{\mu_1 - \mu_m} \leq \frac{\sum_{d_{t,t'} < 0} |d_{t,t'}|}{\sum_{d_{t,t'} > 0} d_{t,t'}}$, which holds for all almost-uniform patterns, following Lemma 4.15 and the definition of almost-uniform patterns.

Lemma 4.15. *The (bounded) PML distribution of any non-trivial pattern satisfies*

$$\frac{\hat{p}_1}{\hat{p}_k} \le \frac{\mu_1 - 1}{\mu_m - 1},$$

where equality holds only if $\mu_t \in \{2, 3\}$ for all $t \in [m]$, and there exists no other (bounded) PML distribution with smaller support size.

If $\mu_1 > 1$ and $\mu_m = 1$, we consider $\frac{\hat{p}_1}{\hat{p}_k} \le \frac{\mu_1 - 1}{\mu_m - 1} = \infty$ to be true.

Lemma 4.16. *For the bounded PML distribution of any pattern, if $\hat{p}_1 > \hat{p}_k$, then*

$$\left(\frac{\hat{p}_1}{\hat{p}_k}\right)^{\mu_1 - \mu_m} \ge \frac{\sum_{d_{t,t'} < 0} |d_{t,t'}|}{\sum_{d_{t,t'} > 0} d_{t,t'}},$$

where the equality holds only if the pattern is 1-uniform and there exists no other bounded PML distribution with smaller support size.

If $d_{t,t} = 0$ for all $t, t' \in [m]$, we consider $\left(\frac{\hat{p}_1}{\hat{p}_k}\right)^{\mu_1 - \mu_m} \ge \frac{0}{0}$ to be true.

Lemma 4.15: Upper Bound for $\frac{\hat{p}_1}{\hat{p}_k}$

Proof of Lemma 4.15. For simplicity, let $P = (p_1, p_2, \ldots, p_k)$, either a PML distribution or a bounded PML distribution. If $\mu_m = 1$, then $\frac{p_1}{p_k} \le \frac{\mu_1 - 1}{\mu_m - 1}$ is trivially true. Without loss of generality suppose $\mu_m > 1$.

If $p_1 = p_k$ then

$$\frac{p_1}{p_k} = 1 < \frac{\mu_1 - 1}{\mu_m - 1}.$$

If $p_1 \neq p_k$, by Proposition 4.14,

$$\sum_{\{t,t'\} \in \binom{[m]}{2}} \Big[(\mu_t - 1)\left(p_1^{\mu_t - 1} p_k^{\mu_{t'}} - p_1^{\mu_{t'}} p_k^{\mu_t - 1}\right) + (\mu_{t'} - 1)\left(p_1^{\mu_{t'} - 1} p_k^{\mu_t} - p_1^{\mu_t} p_k^{\mu_{t'} - 1}\right)\Big] P_{1,k}(\bar{\psi}_{t,t'}) \ge 0,$$

where the equality holds only if $\mu_t \le 3$ for all $t \in [m]$, and there is no other bounded PML distribution with smaller support size.

Then there exists $\{t, t'\} \in \binom{[m]}{2}$ such that

$$(\mu_t - 1)(p_1^{\mu_t - 1} p_k^{\mu_{t'}} - p_1^{\mu_{t'}} p_k^{\mu_t - 1}) \ge (\mu_{t'} - 1)(p_1^{\mu_t} p_k^{\mu_{t'} - 1} - p_1^{\mu_{t'} - 1} p_k^{\mu_t}),$$

where strict inequality can be achieved if $\mu_{t''} > 3$ for some $t'' \in [m]$. Let $r = \frac{p_1}{p_k}$. Dividing both sides of the above inequality by $p_k^{\mu_t+\mu_{t'}-1}$, we get,

$$(\mu_t - 1)(r^{\mu_t-1} - r^{\mu_{t'}}) \geq (\mu_{t'} - 1)(r^{\mu_t} - r^{\mu_{t'}-1}).$$

Then

$$r^{\mu_t} \leq \frac{\mu_t - 1}{\mu_{t'} - 1} r^{\mu_t-1} - \left(\frac{\mu_t - 1}{\mu_{t'} - 1} r - 1\right) r^{\mu_{t'}-1} \leq \frac{\mu_t - 1}{\mu_{t'} - 1} r^{\mu_t-1}.$$

Hence

$$\frac{p_1}{p_k} = r \leq \frac{\mu_t - 1}{\mu_{t'} - 1} \leq \frac{\mu_1 - 1}{\mu_m - 1}. \qquad \square$$

Lemma 4.16: Lower Bound for $\frac{\hat{p}_1}{\hat{p}_k}$

Proof of Lemma 4.16. For simplicity, let $P = (p_1, p_2, \ldots, p_k)$ be the bounded PML distribution $\hat{P}_{\bar{\psi}}^{(K)}$. By Proposition 4.14,

$$\sum_{\{t,t'\} \in \binom{[m]}{2}} c_{t,t'} P_{1,k}(\bar{\psi}_{t,t'}) \geq 0,$$

where

$$c_{t,t'} = (\mu_t - 1)\frac{p_1^{\mu_t-1} p_k^{\mu_{t'}} - p_1^{\mu_{t'}} p_k^{\mu_t-1}}{p_1 - p_k} + (\mu_{t'} - 1)\frac{p_1^{\mu_{t'}-1} p_k^{\mu_t} - p_1^{\mu_t} p_k^{\mu_{t'}-1}}{p_1 - p_k},$$

and the strict inequality holds if there is no other bounded PML distribution with smaller support size.

Without loss of generality, assume $\mu_t \geq \mu'_{t'}$. Let $\delta = \delta_{t,t'} \stackrel{\text{def}}{=} \mu_t - \mu_{t'} \geq 0$. Then

$$\frac{c_{t,t'}}{(p_1 p_k)^{\mu_{t'}-1}} = (\mu_t - 1)(p_1 p_k)\frac{p_1^{\delta-1} - p_k^{\delta-1}}{p_1 - p_k} - (\mu_{t'} - 1)\frac{p_1^{\delta+1} - p_k^{\delta+1}}{p_1 - p_k}.$$

Since

$$\frac{p_1^{\delta+1} - p_k^{\delta+1}}{p_1 - p_k} = (p_1^\delta + p_k^\delta) + p_1 p_k \sum_{t=0}^{\delta-2} p_1^t p_k^{\delta-2-t} = (p_1^\delta + p_k^\delta) + p_1 p_k \frac{p_1^{\delta-1} - p_k^{\delta-1}}{p_1 - p_k}.$$

It follows that

$$\frac{c_{t,t'}}{(p_1 p_k)^{\mu_{t'}-1}} = \delta \cdot \frac{p_1^{\delta+1} - p_k^{\delta+1}}{p_1 - p_k} - (\mu_t - 1)(p_1^\delta + p_k^\delta). \tag{4.19}$$

On the other hand, By Lemma 4.11,

$$\frac{p_1^{\delta+1} - p_k^{\delta+1}}{p_1 - p_k} \leq \frac{1}{2}(\delta+1)(p_1^{\delta} + p_k^{\delta}), \tag{4.20}$$

where the equality holds for all i if and only if $\delta \in \{0, 1\}$. Canceling $\frac{p_1^{\delta+1} - p_k^{\delta+1}}{p_1 - p_k}$ from Equation (4.19) and Inequality (4.20), we get

$$\frac{c_{t,t'}}{(p_1 p_k)^{\mu_{t'}-1}} \leq \left[\frac{1}{2}\delta(\delta+1) - (\mu_t - 1)\right](p_1^{\delta} + p_k^{\delta}),$$

i.e.,

$$\begin{aligned} c_{t,t'} &\leq \frac{1}{2}\left[\delta_{t,t'}^2 - (\mu_t + \mu_{t'} - 2)\right](p_1^{\delta_{t,t'}} + p_k^{\delta_{t,t'}})(p_1 p_k)^{\mu_{t'}-1} \\ &= \frac{1}{2} d_{t,t'}\,(p_1^{\mu_t} p_k^{\mu_{t'}} + p_1^{\mu_{t'}} p_k^{\mu_t})(p_1 p_k)^{-1}. \end{aligned}$$

Thus

$$\sum_{\{t,t'\} \in \binom{[m]}{2}} d_{t,t'}(p_1^{\mu_t} p_k^{\mu_{t'}} + p_1^{\mu_{t'}} p_k^{\mu_t}) P_{1,k}(\bar{\psi}_{t,t'}) \geq 2 \sum_{\{t,t'\} \in \binom{[m]}{2}} c_{t,t'} P_{1,k}(\bar{\psi}_{t,t'}) \geq 0, \tag{4.21}$$

where the second inequality follows from Proposition 4.14, and the equalities hold only if $\delta_{t,t'} \leq 1$ for all pairs $\{t, t'\}$. Let

$$P(1 \to t, k \to t') \stackrel{\text{def}}{=} p_1^{\mu_t} p_k^{\mu_{t'}} P_{1,k}(\bar{\psi}_{t,t'}),$$

the probability that the indices of p_1 and p_k are t and t' respectively. Let

$$P(\{1,k\} \to \{t,t'\}) \stackrel{\text{def}}{=} P(1 \to t, k \to t') + P(1 \to t', k \to t).$$

Then Inequality (4.21) can be rewritten as

$$\sum_{\{t,t'\}:\, t \neq t'} d_{t,t'} P(\{1,k\} \to \{t,t'\}) \geq 0.$$

Let $r = \frac{p_1}{p_k}$. Then

$$P(\{1,k\} \to \{t,t'\}) = \left(1 + r^{-\delta_{t,t'}}\right) P(1 \to t, k \to t'),$$

and hence

$$\sum_{d_{t,t'} > 0} d_{t,t'} \left(1 + r^{-\delta_{t,t'}}\right) P(1 \to t, k \to t') \geq \sum_{d_{t,t'} < 0} |d_{t,t'}| \left(1 + r^{-\delta_{t,t'}}\right) P(1 \to t, k \to t').$$

Note that, for any $i_1, i_2 \in [k]$,

$$\frac{p_{i_1}^{\mu_1} p_1^{\mu_t} p_k^{\mu_{t'}} p_{i_2}^{\mu_m}}{p_1^{\mu_1} p_{i_1}^{\mu_t} p_{i_2}^{\mu_{t'}} p_k^{\mu_m}} = \left(\frac{p_{i_1}}{p_1}\right)^{\mu_1-\mu_t} \left(\frac{p_k}{p_{i_2}}\right)^{\mu_{t'}-\mu_m}.$$

Since P is a bounded PML distribution, it does not have a continuous part. Thus

$$\frac{P(1 \to t, k \to t')}{P(1 \to 1, k \to m)} \le \max_{i_1,i_2 \in [k]} \frac{p_{i_1}^{\mu_1} p_1^{\mu_t} p_k^{\mu_{t'}} p_{i_2}^{\mu_m}}{p_1^{\mu_1} p_{i_1}^{\mu_t} p_{i_2}^{\mu_{t'}} p_k^{\mu_m}} \le 1,$$

and

$$\frac{P(1 \to t, k \to t')}{P(1 \to 1, k \to m)} \ge \min_{i_1,i_2 \in [k]} \frac{p_{i_1}^{\mu_1} p_1^{\mu_t} p_k^{\mu_{t'}} p_{i_2}^{\mu_m}}{p_1^{\mu_1} p_{i_1}^{\mu_t} p_{i_2}^{\mu_{t'}} p_k^{\mu_m}} \ge \left(\frac{p_k}{p_1}\right)^{(\mu_1-\mu_m)-(\mu_t-\mu_{t'})}.$$

It follows that

$$\sum_{d_{t,t'}>0} d_{t,t'} \left(1 + r^{-\delta_{t,t'}}\right) \ge \sum_{d_{t,t'}<0} |d_{t,t'}| \left(1 + r^{-\delta_{t,t'}}\right) \cdot r^{(\mu_t-\mu_{t'})-(\mu_1-\mu_m)},$$

i.e.,

$$\left(\frac{p_1}{p_k}\right)^{\mu_1-\mu_m} \ge \frac{\sum_{d_{t,t'}<0} |d_{t,t'}| \left(1 + r^{\delta_{t,t'}}\right)}{\sum_{d_{t,t'}>0} d_{t,t'} \left(1 + r^{-\delta_{t,t'}}\right)} \ge \frac{\sum_{d_{t,t'}<0} |d_{t,t'}|}{\sum_{d_{t,t'}>0} d_{t,t'}},$$

where the equality holds only if $\delta_{t,t'} \le 1$ for all pairs $\{t, t'\}$. □

Proof for Almost-uniform Patterns

At last, we prove Theorem 4.12 using Lemmas 4.15 and 4.16.

Proof of Theorem 4.12. Suppose for some $K \ge m$, the bounded PML distribution $\hat{P}_{\bar{\psi}}^{(K)} = (p_1, p_2, \ldots, p_k)$ of an almost-uniform pattern $\bar{\psi}$ is not uniform, *i.e.*, $p_1 > p_k$. Furthermore, there are more than one such distributions, choose one that has the smallest support size. Then Lemmas 4.15 and 4.16 imply that,

$$\frac{\sum_{d_{t,t'}<0} |d_{t,t'}|}{\sum_{d_{t,t'}>0} d_{t,t'}} < \left(\frac{p_1}{p_k}\right)^{\mu_1-\mu_m} < \left(\frac{\mu_1 - 1}{\mu_m - 1}\right)^{\mu_1-\mu_m},$$

where the inequalities are strict. On other hand, by the definition of almost-uniform patterns,

$$\frac{\sum_{d_{t,t'}<0} |d_{t,t'}|}{\sum_{d_{t,t'}>0} d_{t,t'}} \ge \left(\frac{\mu_1 - 1}{\mu_m - 1}\right)^{\mu_1-\mu_m},$$

a contradiction. Thus the bounded PML distribution with the smallest support size must be uniform. □

4.4.2 Theorem 4.13: Patterns with Small Multiplicities

For patterns with multiplicities at most 3 Theorem 4.13 provides a relaxed condition for which the PML distribution is uniform, in the sense that such patterns need not to be almost-uniform.

Proof of Theorem 4.13. Suppose for some $K \geq m$, the bounded PML distribution $\hat{P}_{\bar{\psi}}^{(K)} = (p_1, p_2, \ldots, p_k)$ of an almost-uniform pattern $\bar{\psi}$ is not uniform, *i.e.*, $p_1 > p_k$. Furthermore, there are more than one such distributions, choose one that has the smallest support size. By Proposition 4.14,

$$\sum_{\{t,t'\} \in \binom{[m]}{2}} c_{t,t'} P_{i,j}(\bar{\psi}_{t,t'}) > 0, \tag{4.22}$$

where

$$c_{t,t'} = (\mu_t - 1)\frac{p_i^{\mu_t - 1} p_j^{\mu_{t'}} - p_i^{\mu_{t'}} p_j^{\mu_t - 1}}{p_i - p_j} + (\mu_{t'} - 1)\frac{p_i^{\mu_{t'} - 1} p_j^{\mu_t} - p_i^{\mu_t} p_j^{\mu_{t'} - 1}}{p_i - p_j}.$$

In the proof of Lemma 4.16, it has been shown in Equation (4.19) that

$$\frac{c_{t,t'}}{(p_i p_j)^{\mu_{t'} - 1}} = \delta \frac{p_i^{\delta+1} - p_j^{\delta+1}}{p_i - p_j} - (\mu_t - 1)(p_i^\delta + p_j^\delta).$$

For $\mu_t = \mu_{t'} = \mu$,

$$c_{t,t'} = -2(\mu - 1)(p_i p_j)^{\mu - 1}.$$

For $\delta = \mu_t - \mu_{t'} \geq 1$,

$$\begin{aligned}
c_{t,t'} &= (p_i p_j)^{\mu_{t'} - 1} \left[\delta(p_i^{\delta - 1} p_j + \cdots + p_i p_j^{\delta - 1}) - (\mu_{t'} - 1)(p_i^\delta + p_j^\delta)\right] \\
&\leq (p_i p_j)^{\mu_{t'} - 1} \left[\binom{\delta}{2}(p_i^{\delta - 1} p_j + p_i p_j^{\delta - 1}) - (\mu_{t'} - 1)(p_i^\delta + p_j^\delta)\right] \\
&\leq (p_i p_j)^{\mu_{t'} - 1} \cdot \frac{1}{2} d_{t,t'}(p_i^{\delta - 1} p_j + p_i p_j^{\delta - 1}) \\
&= \frac{1}{2} d_{t,t'}(p_i^{\mu_t - 2} p_j^{\mu_{t'}} + p_i^{\mu_{t'}} p_j^{\mu_t - 2}).
\end{aligned}$$

where all equalities hold if and only if $\delta = 1$. Then, by ignoring terms in Equation (4.22) with $\mu_t \neq \mu_{t'}$ such that $d_{t,t'} < 0$, we get

$$\begin{aligned}
\sum_{\mu:\, \varphi_\mu \geq 2} \binom{\varphi_\mu}{2} \cdot 4(\mu - 1)(p_i p_j)^{\mu - 1} P_{i,j}(\bar{\psi}_{r_\mu - 1, r_\mu}) \\
< \sum_{\{t,t'\}:\, d_{t,t'} > 0} d_{t,t'}(p_i^{\mu_t - 2} p_j^{\mu_{t'}} + p_i^{\mu_{t'}} p_j^{\mu_t - 2}) P_{i,j}(\bar{\psi}_{t,t'}),
\end{aligned}$$

where $r_\mu = m - \sum_{\mu' \le \mu} \varphi_{\mu'}$ is the number of multiplicities at most μ, and thus $\bar{\psi}_{r_\mu - 1, r_\mu}$ denotes the pattern with two multiplicities μ removed from $\bar{\psi}$. For patterns with $\mu_1 = 3$, we get

$$4\binom{\varphi_2}{2} P_{i,j}(\bar{\psi}_{r_2-1,r_2}) < \varphi_1 \varphi_3 P_{i,j}(\bar{\psi}_{r_3,m}),$$

where $\bar{\psi}_{r_2-1,r_2}$ is obtained from $\bar{\psi}$ by removing two multiplicities 2, and $\bar{\psi}_{r_3,m}$ is obtained from $\bar{\psi}$ by removing one multiplicity 1 and the other multiplicity 3. In the statement of the theorem, we assumed that $4\binom{\varphi_2}{3} \ge \varphi_1 \varphi_3$. Then

$$P_{i,j}(\bar{\psi}_{r_2-1,r_2}) < P_{i,j}(\bar{\psi}_{r_3,m}).$$

On the other hand, it is easy to see that $\bar{\psi}_{r_2-1,r_2}$ and $\bar{\psi}_{r_3,m}$ have the same length $n-4$ and number of distinct symbols $m-2$, and that $\bar{\psi}_{r_2-1,r_2} \succeq \bar{\psi}_{r_3,1}$. Then by the majorization property in Fact 2.4 we have

$$P_{i,j}(\bar{\psi}_{r_2-1,r_2}) \ge P_{i,j}(\bar{\psi}_{r_3,m}),$$

a contradiction. Therefore our assumption that $p_1 > p_k$ is false; the bounded distribution $P = \hat{P}^{(K)}$ must be uniform. □

Acknowledgment

Section 4.2 appeared partially in The maximum likelihood probability of skewed patterns, Alon Orlitsky and Shengjun Pan. Sections 4.3 and 4.4 appeared partially in *IEEE International Symposium on Information Theory*, 2009. Recent results on pattern maximum likelihood, Jayadev Acharya, Alon Orlitsky and Shengjun Pan, *IEEE Information Theory Worshop, Volos*, Greece, 2009.

Chapter 5

Deterministic Calculation of Pattern Probabilities

In this chapter we consider the calculation of pattern probabilities for any given mixture distribution, not necessarily the PML distribution.

- In Section 5.1 we analyze the complexity of a recursive algorithm.
- In Section 5.2 we represent the pattern probability using power sums.
- In Section 5.3 we present proofs for a related graph theory problem.

5.1 Recursive Algorithm

Recall that a pattern $\bar{\psi}$ can be regarded as a set. It is easy to see that, as sets,

$$\bar{\psi} = 1^{\mu_1} \times 1^{\mu_2} \cdots (m-1)^{\mu_m} \setminus \bigcup_{i \geq 2} \bar{\psi}^{(i)},$$

where $\bar{\psi}^{(i)}$ is the canonical pattern with multiplicities

$$\mathcal{M}(\bar{\psi}^{(i)}) = \{\mu_1 + \mu_i, \mu_2, \ldots, \mu_{i-1}, \mu_{i+1}, \mu_m\}^* .$$

Thus for any distribution $P = (p_1, p_2, \ldots, p_k) \in \mathcal{P}_{\text{mix}}^{\text{sorted}}$,

$$P(\bar{\psi}) = P(1^{\mu_1}) \cdot P\Big(1^{\mu_2} \cdots (m-1)^{\mu_m}\Big) - \sum_{i=2}^{m} P(\bar{\psi}^{(i)}). \tag{5.1}$$

For example, for $\bar{\psi} = 1^4 2^2 3^2$, we have $\bar{\psi}^{(2)} = \bar{\psi}^{(3)} = 1^6 2^2$, and

$$P(1^4 2^2 3^2) = P(1^4) \cdot P(1^2 2^2) - P(1^6 2^2) - P(1^6 2^2).$$

For simplicity, for any list of numeric (real or integral) values $L_1, L_2, \ldots$ and a set of indices $S \subseteq \{1, 2, \ldots\}$, let

$$L_S = \sum_{i \in S} L_i.$$

For example, $\mu_S = \sum_{i \in [m]} \mu_i$ and $p_I = \sum_{i \in I} p_i$.

Observe that all patterns of the form 1^μ appearing in the recursive calculation satisfy $\mu = \mu_S$ for some $S \subseteq [m]$. Given pattern $\bar{\psi}$ and distribution P, Equation (5.1) gives a recursive algorithm to calculate $P(\bar{\psi})$:

Step 1: Calculate $P(1^{\mu_S})$ for all $S \subseteq [m]$.

Step 2: Recursively calculate $P(\bar{\psi})$ using Equation (5.1).

We show the following arithmetic computational complexities of the algorithm.

Theorem 5.1. *For any distribution P, the probability of pattern $\bar{\psi} = 1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m}$ can be deterministically calculated in time*

$$O\big(k \min\{n, 2^m\} \log n + e^{\pi\sqrt{\frac{2n}{3}}} \log m\big).$$

Theorem 5.2. *For any distribution P, the probability of pattern $\bar{\psi} = 1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m}$ can be deterministically calculated in time*

$$O\big(k \min\{n, 2^m\} \log n + \min\{n2^m, 3^m, nm^d, m^{2d}\} \log m\big).$$

5.1.1 Complexity for Step 1

In Step 1 $P(1^{\mu_S}) = \sum_{i=1}^{k} p_i^{\mu_S}$ can be calculated in time $O(k \log \mu_S)$ by repeatedly taking squares of p_i all $i \in [k]$ then sum up the final powers $p_i^{\mu_S}$. Note that $\mu_S \leq n$, and the number of distinct values of μ_S is at most $\min\{n, 2^m\}$. Then Step 1 can be done in time $O(k \min\{n, 2^m\} \log n)$. The big-O notation here refers to arithmetic complexity, not bit-complexity.

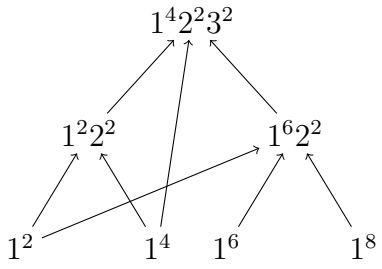


Figure 5.1: Computation Graph for Pattern $\bar{\psi} = 1^4 2^2 3^2$

5.1.2 Theorem 5.1: First Complexity for Step 2

Note that in Step 2 the same pattern may appear multiple times on the right-hand of Equation (5.1). To avoid unnecessary repetition, we define the *computation graph*$G(\bar{\psi})$ for a pattern $\bar{\psi}$, which is a directed graph, consisting of canonical forms of all patterns ever appearing in Step 2. The (outgoing) neighbors of a pattern $\bar{\psi}' = 1^{\mu'_1} 2^{\mu'_2} \cdots m^{\mu'_m}$ are patterns on the right-hand side of Equation (5.1), including

(1) $1^{\mu'_1}$,

(2) $1^{\mu'_2} \cdots (m'-1)^{\mu'_m}$, and

(3) $\bar{\psi}'^{(i)}$ for all $i \in [2\,..\,m']$.

For example, for pattern $\bar{\psi} = 1^4 2^2 3^2$, $G(\bar{\psi})$ is illustrated in Figure 5.1.

Proof of Theorem 5.1. We just need to show that Step 2 can be done in time $O(me^{\pi\sqrt{\frac{2n}{3}}})$. We first consider the number of nodes (patterns) in the computation graph $G(\bar{\psi})$. It is easy to see that the multiplicities of any pattern in the computation graph can be obtained by first take a subset of $\mathcal{M}(\bar{\psi})$, then partition it into disjoint subsets and sum the multiplicities in each subset.

Recall that the nodes in $G(\bar{\psi})$ are all canonical. It is easy to see that given n, there is a one-to-one correspondence between canonical patterns and partitions of n into unordered positive numbers (the multiplicities). Let $p(n)$ be the number of such partitions. Then the number of nodes in $T(\bar{\psi})$ is at most $\sum_{i=1}^{n} p(n)$.

It is know that (*e.g.* [vLW92]), asymptotically

$$p(n) \sim \frac{1}{4\sqrt{3}\,n} e^{\pi\sqrt{\frac{2n}{3}}}.$$

Thus the number of nodes in $G(\bar{\psi})$ is $O\left(e^{\pi\sqrt{\frac{2n}{3}}}\right)$, and it follows that Step 2 can be done in time

$$O\left(e^{\pi\sqrt{\frac{2n}{3}}}\,\log m\right),$$

where the factor $O(\log m)$ is the cost for calculating the product and sum on the right-hand side of Recursion (5.1), assuming the pattern probabilities on the right-hand side are already calculated. □

5.1.3 Theorem 5.2: Second Complexity for Step 2

Recall that the computation graph $G(\bar{\psi})$ consists of canonical patterns that appear in Step 2 of calculating $P(\bar{\psi})$ using the recursion (5.1).

We can further reduce unnecessary computation by the following observation. The multiplicities of any pattern in the computation graph can be obtained by first partition a subset of $\mathcal{M}(\bar{\psi})$ into disjoint subsets, and then sum the multiplicities in each subset. However, note that in Equation (5.1) we always merge the largest multiplicity with another multiplicity. Thus any pattern in the computation graph can actually be obtained as follows:

(1) Partition $\mathcal{M}(\bar{\psi})$ into three disjoint sub-multisets

$$\mathcal{M}_0,\, \mathcal{M}_1,\, \mathcal{M}_2,$$

where $\mathcal{M}_2$ could be empty while the other two are not.

(2) Form a canonical pattern with multiplicities

$$\Big\{ \sum_{\mu\in\mathcal{M}_0} \mu \Big\} \bigcup \mathcal{M}_1,$$

where the union is for multisets.

Let $\nu_1 > \nu_2 > \cdots > \nu_d > 0$ be the distinct multiplicities in a pattern $\bar{\psi}$, where d is the number of distinct multiplicities. Let $\varphi_1^+, \varphi_2^+, \ldots, \varphi_d^+$ be the corresponding prevalences, where the superscript + denotes that these prevalences are positive.

For example, for $\bar{\psi} = 1^3 2^2 3^2 4567$, the distinct multiplicities are $\nu_1 = 3, \nu_2 = 2, \nu_3 = 1$, and the corresponding prevalenes are $\varphi_1^+ = 1, \varphi_2^+ = 2, \varphi_3^+ = 4$.

Then a pattern $\bar{\psi}'$ in $G(\bar{\psi})$ can be represented by its *profile-form*

$$\Big(\sum_{i=1}^{d} \phi_i' \nu_i; \phi_1, \phi_2, \ldots, \phi_d \Big),$$

where ϕ_i and ϕ_i' are the number of multiplicities μ in $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively. In addition, ϕ_μ and ϕ_μ' could be 0, and

$$\phi_\mu + \phi_\mu' \leq \varphi_\mu'.$$

For example, for $\bar{\psi} = 1^3 2^2 3^2 4567$, the distinct multiplicities are $\nu_1 = 3, \nu_2 = 2, \nu_3 = 1$, and the corresponding prevalenes are $\varphi_1^+ = 1, \varphi_2^+ = 2, \varphi_3^+ = 4$. Pattern $1^5 23$ appears in the computation graph, corresponding to the partition $\mathcal{M}_0 = \{3, 2\}, \mathcal{M}_1 = \{1, 1\}, \mathcal{M}_2 = \{2, 1, 1\}$, and the profile-form of $1^5 23$ is $(5; 0, 0, 2)$, that is, we merge the multiplicities 3 and 2, and then select no multiplicities 3 or 2, but select 2 multiplicities 1.

Proof of Theorem 5.2. We will find an upper bound for the number of patterns in $G(\bar{\psi})$ in an approach different that in the proof of Theorem 5.1. Recall that a pattern in $G(\bar{\psi})$ can be obtained from a partition of $\mathcal{M}(\bar{\psi})$ into three parts $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$, then removing multiplicities in $\mathcal{M}_1$ and merge multiplicities in $\mathcal{M}_3$, and it can be represented by its profile-form

$$\Big(\sum_{i=1}^{d} \phi_i' \nu_i; \phi_1, \phi_2, \ldots, \phi_d \Big),$$

Thus we only need to count the number of profile-forms.

Note that $\phi_i' \geq 0$, $\phi_i \geq 0$, and $\phi_i' + \phi_i \leq \varphi_i^+$. It can be shown that the number of possible pairs (ϕ_i', ϕ_i) is exactly $\binom{\varphi_i^+ + 1}{2}$. Thus the number of profile-forms is at most

$$\binom{\varphi_i^+ + 1}{2}.$$

On the other hand, note that $\sum_{i=1}^{d} \phi_i' \nu_i < n$ and $\phi_i \leq \varphi_i^+$. Then the number of profile-forms can also be bounded as

$$n \prod_{i=1}^{d} (\varphi_i^+ + 1).$$

Thus the computational cost for Step 2 is

$$O\left(\log m \cdot \min\left\{ \prod_{i=1}^{d} \binom{\varphi_i^+ + 2}{2}, n \prod_{i=1}^{d} (\varphi_i^+ + 1) \right\} \right),$$

where $\log m$ is again the cost for calculating the product and sum on the right-hand side of the Equation (5.1). We simplify the bounds as follows. Since $\sum_{i=1}^{d} \varphi_i^+ = m$,

$$\prod_{i=1}^{d} (\varphi_i^+ + 1) \leq \left(\frac{\sum_{i=1}^{d} (\varphi_i + 1)}{d} \right)^d = \left(\frac{m}{d} + 1 \right)^d \leq 2^m,$$

where the last inequality follows from $d \leq m$. Similarly,

$$\prod_{i=1}^{d} \frac{\varphi_i^+ + 2}{2} \leq \left(\frac{m}{2d} + 1 \right)^d \leq 1.5^m.$$

It follows that

$$\log m \, n \prod_{i=1}^{d} (\varphi_i^+ + 1) \leq n 2^m \log m, \text{ and}$$

$$\log m \prod_{i=1}^{d} \binom{\varphi_i^+ + 2}{2} = \log m \prod_{i=1}^{d} (\varphi_i^+ + 1) \prod_{i=1}^{d} \frac{\varphi_i^+ + 2}{2} \leq 3^m \log m.$$

This proves two of bounds in the conclusion. Note that

$$\left(\frac{m}{d} + 1 \right)^d \leq m^d, \text{ and } \left(\frac{m}{2d} + 1 \right)^d \leq m^d.$$

Then the other two bounds follow. □

5.2 Formulation in Power sums

The algorithm using Recursion (5.1) achieves efficiency in time at the cost of memory storage. Given pattern $\bar{\psi} = 1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m}$ and distribution $P =$

$(p_1, p_2, \ldots, p_k)$, a direct calculation of $P(\bar{\psi})$ as sum of sequence probabilities requires constant memory space. However, the computational time is in the order of the size of $\bar{\psi}$, i.e., $k^{\underline{m}}$.

We show that the pattern probabilities can be calculated more efficiently in memory cost, but with time complexity linear in the support size:

$$O(k \min\{n, 2^m\} \log n + m^m \log m).$$

The idea is to write as a polynomial in power sums

$$P(1^t) = \sum_{i=1}^{k} p_i^t, \quad 2 \leq t \leq n.$$

Recall that $\mathcal{M}(\bar{\psi}) = \{\mu_1, \mu_2, \ldots, \mu_m\}$ is the multiset of multiplicities of pattern $\bar{\psi}$. Let $\mathcal{G}_{\bar{\psi}}$ be the set of all graphs over $\mathcal{M}(\bar{\psi})$. Given $G \in \mathcal{G}_{\bar{\psi}}$, let $V_1, V_2, \ldots, V_D$ be the vertex sets of the connected components of G. Then G *induces* a disjoint partition of $\mathcal{M}(\bar{\psi})$, denoted as

$$\mathbb{P}_G \stackrel{\text{def}}{=} \{V_1, V_2, \ldots, V_D\}.$$

The *probability of graph* G is

$$\Pr(G) \stackrel{\text{def}}{=} \prod_{i=1}^{t} P(1^{\mu_{V_i}}),$$

where $\mu_{V_i} = \sum_{t \in V_i} \mu_t$.

For example, consider the pattern $\bar{\psi} = 1^{\mu_1} 2^{\mu_2} \cdots 5^{\mu_5}$. Let G be the graph with edges

$$E = \{\{\mu_1, \mu_2\}, \{\mu_2, \mu_3\}, \{\mu_4, \mu_5\}\}.$$

Then G has two connected components with vertex sets

$$V_1 = \{\mu_1, \mu_2, \mu_3\}, \text{ and } V_2 = \{\mu_4, \mu_5\}.$$

Then

$$\Pr(G) = P(1^{\mu_1+\mu_2+\mu_3}) P(1^{\mu_4+\mu_5}).$$

An *even*(*odd*) graph has even (odd) number of edges. Let

$$\texttt{sign}\,(G) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } G \text{ is even} \\ -1, & \text{if } G \text{ is odd,} \end{cases}$$

i.e., $\texttt{sign}\,(G) = (-1)^{|E(G)|}$.

We show that the pattern probability $P(\bar{\psi})$ can be written as a summation over graphs in $\mathcal{G}_{\bar{\psi}}$.

Theorem 5.3. *For any pattern $\bar{\psi}$ and distribution P,*

$$P(\bar{\psi}) = \sum_{G \in \mathcal{G}_{\bar{\psi}}} \texttt{sign}\,(G)\Pr(G). \tag{5.2}$$

Theorem 5.3 can be used to calculate $P(\bar{\psi})$ as follows. First we pre-compute all power sums $P(1^{\mu_S})$ for all $S \subseteq [m]$, which we have shown previously to take time $O(k \min\{n, 2^m\} \log n)$. Then for each graph $G \in G_{\bar{\psi}}$, we calculate $\Pr(G)$ as a product of at most m power sums. Since there are $2^{\binom{m}{2}}$ graphs, the overall complexity is

$$O\left(k \min\{n, 2^m\} \log n + 2^{\binom{m}{2}} \log m\right).$$

Note that any graph $G \in \mathcal{G}_{\bar{\psi}}$ induces a partition $\mathbb{P}_G = \{V_1, V_2, \ldots, V_d\}$ of $\mathcal{M}(\bar{\psi})$. Note that the V_i's are sufficient for calculating $\Pr(G)$. Let $\mathcal{P}_{\bar{\psi}}$ be the set of all partitions of $\mathcal{M}(\bar{\psi})$. The *probability of a partition* $\mathbb{P} \in \mathcal{P}_{\bar{\psi}}$, denoted by $\Pr(\mathbb{P})$, is the probability of any graph whose vertex sets of components are the parts in $\mathbb{P}$. In other words,

$$\Pr(\mathbb{P}) = \prod_{V \in \mathbb{P}} P(1^{\mu_V}),$$

where the product is over all subsets C of $\mathcal{M}(\bar{\psi})$ in $\mathbb{P}$. Thus in the formulation from Theorem 5.3 we may group together graphs that induce the partition. Given partition $\mathbb{P} = \{V_1, V_2, \ldots, V_{|\mathbb{P}|}\} \in \mathcal{P}_{\bar{\psi}}$, where $|\mathbb{P}|$ is the number of subsets in $\mathbb{P}$, let $n_i = |V_i|$ be the size of V_i. We will show that the pattern probability can also be written as a summation over partitions:

Theorem 5.4. *For any pattern $\bar{\psi}$ and distribution P,*

$$P(\bar{\psi}) = \sum_{\mathbb{P} \in \mathcal{P}_{\bar{\psi}}} \prod_{i=1}^{|\mathbb{P}|} (-1)^{n_i - 1} (n_i - 1)!\, P(1^{n_i}).$$

Similarly, using Theorem 5.4 we can calculate $P(\bar{\psi})$ using Theorem 5.4 by first pre-computing all power sums $P(1^{n'})$ as well as factorials $n'!$ for $n' \in [n]$, which takes time $O(k \min\{n, 2^m\} \log n)$. Here we refer $O(\cdot)$ to the arithmetic complexity. For each partition $\mathbb{P}$ the computation of $\prod_{i=1}^{|\mathbb{P}|} (-1)^{n_i - 1} (n_i - 1)!\, P(1^{n_i})$ can be done in time $O(\log |\mathbb{P}|) = O(\log m)$. The total number of partitions is at most m^m. Thus the overall complexity is

$$O\left(k \min\{n, 2^m\} \log n + m^m \log m\right).$$

Further improvement on calculating $P(\bar{\psi})$ can be obtained by considering profiles. Note that any partition $\mathbb{P} \in \mathcal{P}_{\bar{\psi}}$ *induces* a decomposition of the pattern $\bar{\psi}$ into shorter patterns. Recall that the profile of a pattern is the list of prevalences:

$$\bar{\varphi} \stackrel{\text{def}}{=} (\varphi_1, \varphi_2, \ldots, \varphi_{\mu_{\max}}),$$

where $\mu_{\max} = \mu_1$ is the largest multiplicity. Let the *multi-profile* of a partition $\mathbb{P} \in \mathscr{P}_{\bar{\psi}}$, denoted by $\boldsymbol{\varphi}_{\mathbb{P}}$, be the multiset of profiles of the shorter patterns decomposed by $\mathbb{P}$.

For example, for $\bar{\psi} = 1^5 2^3 3^3 4^3$, the partition $\mathbb{P} = \{\{5, 3\}, \{3, 3\}\}$, decomposes $\bar{\psi}$ into two shorter patterns $1^5 2^3$ and $1^3 2^3$, which have profiles $(0, 0, 1, 0, 1)$ and $(0, 0, 2)$ respectively. Then the multi-profile of $\mathbb{P}$ is $\boldsymbol{\varphi} = \{(0, 0, 1, 0, 1), (0, 0, 2)\}$.

For a given pattern $\bar{\psi}$, let $\mathscr{P}_{\boldsymbol{\varphi}}$ be the set of all partitions $\mathbb{P} \in \mathscr{P}_{\bar{\psi}}$ having the same multi-profile $\boldsymbol{\varphi}$. Let $\bar{\varphi}^1, \bar{\varphi}^2, \ldots, \bar{\varphi}^D$ be the distinct profiles in $\boldsymbol{\varphi}$, where D is the total number of distinct profiles. Let d_i be the number of occurrences of $\bar{\varphi}^i$ in $\mathscr{P}_{\boldsymbol{\varphi}}$. Let $\bar{\varphi}^i_\mu$ be the prevalence of μ in $\bar{\varphi}^i$. Then by combinatorial arguments the number of partitions in $\mathscr{P}_{\boldsymbol{\varphi}}$ is

$$|\mathscr{P}_{\boldsymbol{\varphi}}| = \frac{\prod_{\mu > 0} \binom{\varphi_\mu}{\bar{\varphi}^1_\mu, \bar{\varphi}^2_\mu, \ldots, \bar{\varphi}^D_\mu}}{\prod_{i=1}^{D} d_i!}.$$

Note that the multi-profile of a partition is sufficient for calculating its probability. Define the *probability of multi-profile* $\boldsymbol{\varphi}$, denoted by $\Pr(\boldsymbol{\varphi})$, to be the probability

of any partition whose multi-profile is $\boldsymbol{\varphi}$. Then

$$\Pr(\boldsymbol{\varphi}) = \prod_{\bar{\varphi}\in\boldsymbol{\varphi}} P(1^{\sum_{\mu>0}\mu\varphi_\mu}) = \prod_{i=1}^{D} P\left(1^{\sum_\mu \mu\bar{\varphi}^i_\mu}\right)^{d_i}.$$

We can rewrite the formulation of $P(\bar{\psi})$ in Theorem 5.4 by grouping together partitions that have the same multi-profile.

Let Φ be the set of all possible distinct multi-profiles from pattern $\bar{\psi}$. For any profile $\bar{\varphi}$ let $|\bar{\varphi}| \stackrel{\text{def}}{=} \{\mu : \bar{\varphi}_\mu > 0\}$. Then it follows from Theorem 5.4 that

Corollary 5.5. *For any pattern $\bar{\psi}$ and distribution P,*

$$P(\bar{\psi}) = \sum_{\boldsymbol{\varphi}\in\Phi} \frac{\prod_{\mu>0}\binom{\varphi_\mu}{\bar{\varphi}^1_\mu,\bar{\varphi}^2_\mu,\ldots,\bar{\varphi}^D_\mu}}{\prod_{i=1}^D d_i!} \cdot \prod_{i=1}^{D}\left[(-1)^{|\bar{\varphi}^i|-1}(|\bar{\varphi}^i|-1)P(1^{\sum_\mu\mu\bar{\varphi}^i_\mu})\right]^{d_i}.$$

Similar to calculating $P(\bar{\psi})$ using Theorem 5.3 or 5.4, we can calculate $P(\bar{\psi})$ using Corollary 5.5 by first pre-computing all power sums $P(1^{n'})$ as well as factorials $n'!$ for $n' \in [n]$, which takes time $O(k\min\{n, 2^m\}\log n)$. the computation of each term for a multi-profile can be done in time $O(D) = O(\log m)$. Thus the overall complexity is

$$O\left(k\min\{n, 2^m\}\log n + |\Phi|\log m\right),$$

where $|\Phi|$ is the number of all distinct multi-profiles from $\bar{\psi}$, which can be bounded as follows. Let $\bar{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_{\mu_{\max}})$ be the profile of the pattern $\bar{\psi}$. Then a multi-profile $\boldsymbol{\varphi} \in \Phi$ can regarded as an unordered partition of $\bar{\varphi}$ into vectors with nonnegative integer values.

We mentioned that the partition number satisfies $p(n) \sim \frac{1}{4\sqrt{3}\,n} e^{\pi\sqrt{\frac{2n}{3}}}$. Similar to the proof using generating functions in [vLW92], we show a general counting argument for the number of vector partitions.

Lemma 5.6. *Given a vector $\bar{n} = (n_1, n_2, \ldots, n_d)$ of positive integers such that $n = \sum_{i=1}^d n_i \geq 4$, the number of partitions $p(\bar{n})$ of $\bar{n}$ into vectors of nonnegative integers can be bounded as*

$$p(\bar{n}) \leq \exp\left(2n^{\frac{d}{d+1}} + n\ln\left(1 + 2d\,n^{-\frac{1}{d+1}}\right)\right).$$

Note: It was shown in [DO06] that, for sufficiently large n and $d = o(\ln(n))$,

$$\ln p(\bar{n}) \leq \frac{d+1}{d}\left(\frac{\pi^2 d^2}{6}\right)^{\frac{1}{d+1}} n^{\frac{d}{d+1}} + O(n^{\frac{d-1}{d+1}}) \leq 3n^{\frac{d}{d+1}} + O(n^{\frac{d-1}{d+1}}).$$

As a comparison, the bound in Lemma 5.6 holds for all $n \geq 4$ and $d \geq 1$.

The bound in Lemma 5.6 can be simplified as follows:

$$p(\bar{n}) \leq e^{2n^{\frac{d}{d+1}}} \cdot \left(1 + \frac{2d}{n^{\frac{1}{d+1}}}\right)^n \leq e^{2n^{\frac{d}{d+1}}} \cdot e^{2dn^{\frac{d}{d+1}}} = e^{2(d+1)n^{\frac{d}{d+1}}}.$$

Thus we can calculate $P(\bar{\psi})$ using Corollary 5.5 in time

$$O\Big(k \min\{n, 2^m\} \log n + e^{2(d+1)m^{\frac{d}{d+1}}} \log m\Big).$$

5.2.1 Theorem 5.3: Expansion over Graphs

It is easy to see $P(\bar{\psi})$ is a symmetric polynomials in the discrete probabilities $p_1, p_2, \ldots, p_k$. For $t \in [n]$, the polynomials

$$e_t(p_1, p_2, \ldots, p_k) \stackrel{\text{def}}{=} \sum_{I \in \binom{[n]}{t}} \prod_{i \in I} p_i$$

are the *elementary symmetric polynomials*. Clearly for discrete distributions

$$P(12\cdots n) = \frac{e_n}{n!}.$$

The well-known Fundamental Theorem of Symmetric Polynomials states that any symmetric polynomial can be represented as a polynomial of the elementary symmetrical polynomials. Furthermore, for $t \in [2..n]$, the power sums $P(1^t) = \sum_{i=1}^k p_i^t$ are also symmetric polynomials. Newton-Girard Formulae give a way to write an elementary symmetrical polynomial in terms of the power sums. Thus the pattern probability can be written as polynomial in the power sums. Recursion can be used to obtain the power-sum representation. However, we show that such representation can be obtained in a closed form.

Proof of Theorem 5.3. Let $\mathcal{A}$ be the underlying alphabet of P. Let

$$U \stackrel{\text{def}}{=} \{x_1^{\mu_1} x_2^{\mu_2} \cdots x_m^{\mu_m} : x_i \in \mathcal{A}\, \forall i \in [m]\},$$

the set of sequences consisting of runs of lengths $\mu_1, \mu_2, \ldots, \mu_m$. Here x_i's are not necessarily distinct. For any pair $\{i,j\} \in \binom{[m]}{2} = \{(i,j) \mid i,j \in [m], i \neq j\}$, let

$$S_{i,j} \stackrel{\text{def}}{=} \{x_1^{\mu_1} x_2^{\mu_2} \cdots x_m^{\mu_m} \in U : x_i = x_j\}.$$

Then

$$\bar{\psi} = \bigcap_{\{i,j\} \in \binom{[m]}{2}} U \setminus S_{i,j} = U \setminus \bigcup_{\{i,j\} \in \binom{[m]}{2}} S_{i,j}.$$

Thus

$$P(\bar{\psi}) = P(U) - P\Big(\bigcup_{\{i,j\} \in \binom{[m]}{2}} S_{i,j} \Big).$$

Using the inclusion-exclusion principle, we get

$$P(\bar{\psi}) = P(U) + \sum_{I \subseteq \binom{[m]}{2} : I \neq \emptyset} (-1)^{|E|} P\left(\bigcap_{\{i,j\} \in E} S_{i,j} \right). \tag{5.3}$$

Observe that any subset $E \subseteq \binom{[m]}{2}$ uniquely determines a graph G_E over $\mathcal{M}(\bar{\psi})$ with edge set E. For $E \neq \emptyset$,

$$P\Big(\bigcap_{\{i,j\} \in E} S_{i,j} \Big) = \Pr(G_E),$$

and $P(U) = \Pr(G_\emptyset)$. It follows from Equation (5.3) that $P(\bar{\psi})$ can be written as $\sum_{G \in \mathcal{G}_{\bar{\psi}}} \mathtt{sign}\,(G) \Pr(G)$. □

Example: Consider pattern $1^{\mu_1} 2^{\mu_2} 3^{\mu_3}$. As shown in Figure 5.2, there are 8 graphs over the multiset $\mathcal{M}(\bar{\psi}) = \{\mu_1, \mu_2, \mu_3\}^*$: the empty graph, three graphs with one edge, three graphs with two edges, and the complete graph. Then

$$\begin{aligned} P(1^{\mu_1} 2^{\mu_2} 3^{\mu_3}) = {} & P(1^{\mu_1}) P(1^{\mu_2}) P(1^{\mu_3}) - P(1^{\mu_2+\mu_3}) P(1^{\mu_1}) \\ & - P(1^{\mu_1+\mu_3}) P(1^{\mu_2}) - P(1^{\mu_1+\mu_2}) P(1^{\mu_3}) \\ & + 3P(1^{\mu_1+\mu_2+\mu_3}) - P(1^{\mu_1+\mu_2+\mu_3}). \end{aligned}$$

5.2.2 Theorem 5.4: Expansion over Partitions

Given a set of graphs $\mathcal{G}$, let $\mathtt{even}(\mathcal{G})$ and $\mathtt{odd}(\mathcal{G})$ be the number of even and odd graphs in $\mathcal{G}$ respectively. Let

$$\mathtt{diff}(\mathcal{G}) \stackrel{\text{def}}{=} \sum_{G \in \mathcal{G}} \mathtt{sign}\,(G) = \mathtt{even}(\mathcal{G}) - \mathtt{odd}(\mathcal{G}).$$

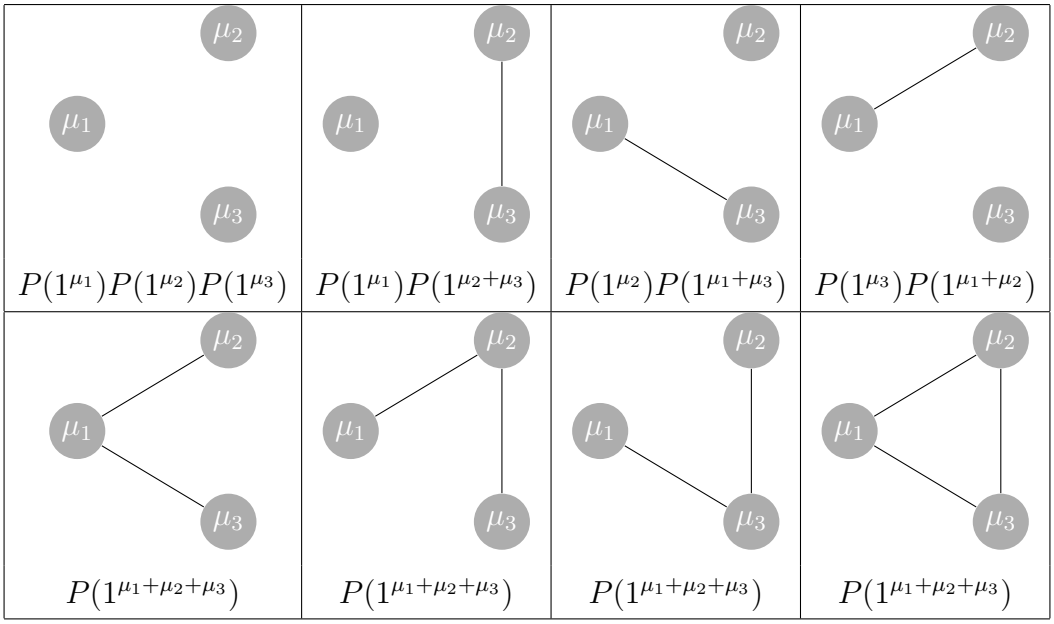


Figure 5.2: Example for Calculating $P(1^{\mu_1}2^{\mu_2}3^{\mu_3})$

It is straightforward to verify the following proprieties for any two disjoint sets $\mathcal{G}_1$ and $\mathcal{G}_2$ of graphs:

(1) If $\mathcal{G}_1 \cap \mathcal{G}_2 = \emptyset$, then

$$\texttt{diff}(\mathcal{G}_1 \cup \mathcal{G}_2) = \texttt{diff}(\mathcal{G}_1) + \texttt{diff}(\mathcal{G}_2).$$

(2) If $E(G_1) \cap E(G_2) = \emptyset$ for any $G_1 \in \mathcal{G}_1$ and $G_2 \in \mathcal{G}_2$,

$$\texttt{diff}(\mathcal{G}_1 \otimes \mathcal{G}_2) = \texttt{diff}(\mathcal{G}_1) \cdot \texttt{diff}(\mathcal{G}_2),$$

where $\mathcal{G}_1 \otimes \mathcal{G}_2 \stackrel{\text{def}}{=} \{G_1 \cup G_2 \mid G_1 \in \mathcal{G}_1, G_2 \in \mathcal{G}_2\}$.

To prove Theorem 5.4, we need the following general counting argument. For any positive integer n, let $\mathcal{C}_n$ be the set of all connected graphs over vertex set $[n]$. Then

Lemma 5.7. *For any $n \geq 1$,*

$$\texttt{diff}(\mathcal{C}_n) = (-1)^{n-1}(n-1)!.$$

We will give several proofs of Lemma 5.7 later in Sectionsec:evenodd. For any partition $\mathbb{P} \in \mathcal{P}_{\bar{\psi}}$, let

$$\mathcal{G}_{\mathbb{P}} \stackrel{\text{def}}{=} \{G \in \mathcal{G}_{\bar{\psi}} \mid \mathbb{P}_G = \mathbb{P}\},$$

the set of all graphs that induce the same partition $\mathbb{P}$.

Proof of Theorem 5.4. We have shown in Theorem 5.3 that

$$P(\bar{\psi}) = \sum_{G \in \mathcal{G}_{\bar{\psi}}} \mathtt{sign}\,(G) \Pr(G).$$

By grouping graphs with the same induced partition of $\mathcal{M}(\bar{\psi})$, we can rewrite $P(\bar{\psi})$ as

$$P(\bar{\psi}) = \sum_{\mathbb{P} \in \mathcal{P}_{\bar{\psi}}} \left[\sum_{G \in \mathcal{G}_{\mathbb{P}}} \mathtt{sign}\,(G) \right] \Pr(\mathbb{P}) = \sum_{\mathbb{P} \in \mathcal{P}_{\bar{\psi}}} \mathtt{diff}(\mathcal{G}_{\mathbb{P}}) \Pr(\mathbb{P}).$$

Given partition $\mathbb{P} = \{V_1, V_2, \ldots, V_{|\mathbb{P}|}\}$, where $|\mathbb{P}|$ is the number of subsets in $\mathbb{P}$, let $\mathcal{C}_i$ be the set of connected graphs over V_i. Then

$$\mathcal{G}_{\mathbb{P}} = \mathcal{C}_1 \otimes \mathcal{C}_2 \otimes \cdots \otimes \mathcal{C}_{|\mathbb{P}|}.$$

It follows from the property of $\mathtt{diff}$ that $\mathtt{diff}(\mathbb{P}) = \prod_{i=1}^{|\mathbb{P}|} \mathtt{diff}(\mathcal{C}_i)$. Note that $\Pr(\mathbb{P}) = \prod_{i=1}^{|\mathbb{P}|} P(1^{n_i})$, where $n_i = |V_i|$. Then by Lemma 5.7

$$P(\bar{\psi}) = \sum_{\mathbb{P} \in \mathcal{P}_{\bar{\psi}}} \prod_{i=1}^{|\mathbb{P}|} \mathtt{diff}(\mathcal{C}_i) \cdot \prod_{i=1}^{|\mathbb{P}|} P(1^{n_i}) = \sum_{\mathbb{P} \in \mathcal{P}_{\bar{\psi}}} \prod_{i=1}^{|\mathbb{P}|} (-1)^{n_i - 1} (n_i - 1)!\, P(1^{n_i}). \qquad \square$$

5.2.3 Lemma 5.7: Expansion over Multi-profiles

Proof of Lemma 5.6. Let $\bar{n} = \sum_{\bar{v} \geq 0} \varphi_{\bar{v}} \bar{v}$ be a partition of $\bar{n}$, where the summation is over distinct vectors $\bar{v}$, and $\varphi_{\bar{v}}$ is the number of $\bar{v}$'s. Then for variables $x_1, x_2, \ldots, x_d$,

$$\prod_{i=1}^{d} x_i^{n_i} = \prod_{\bar{v} \geq 0} \left(\prod_{i=1}^{d} x_i^{v_i} \right)^{\varphi_{\bar{v}}},$$

where we use $\bar{v} \geq 0$ to denote that $v_i \geq 0$ for all $i \in [d]$. Hence

$$p(\bar{n}) = \left[\prod_{i=1}^{d} x_i^{n_i} \right] \prod_{\bar{v} \geq 0} \sum_{\varphi \geq 0} \left(\prod_{i=1}^{d} x_i^{v_i} \right)^{\varphi} = \left[\prod_{i=1}^{d} x_i^{n_i} \right] \prod_{\bar{v} \geq 0} \frac{1}{1 - \prod_{i=1}^{d} x_i^{v_i}},$$

where $\bar{v} \gneq 0$ denotes that $v_i \geq 0$ for all $i \in [d]$ but $v_{i'} > 0$ for some $i' \in [d]$. It follows that the generating function for $p(\bar{n})$ is

$$f(\bar{x}) \stackrel{\text{def}}{=} \sum_{\bar{n} \geq 0} p(\bar{n}) \prod_{i=1}^{d} x_i^{n_i} = \prod_{\bar{n} \gneq 0} \frac{1}{1 - \prod_{i=1}^{d} x_i^{n_i}}.$$

For $\bar{x}$ such that $0 < x_i < 1$ for all $i \in [d]$, we get

$$\begin{aligned}
\ln f(\bar{x}) &= \sum_{\bar{n} \gneq 0} -\ln \left(1 - \prod_{i=1}^{d} x_i^{n_i}\right) \\
&= \sum_{\bar{n} \gneq 0} \sum_{t \geq 1} \frac{1}{t} \left(\prod_{i=1}^{d} x_i^{n_i}\right)^t \\
&= \sum_{t \geq 1} \frac{1}{t} \sum_{\bar{n} \gneq 0} \prod_{i=1}^{d} x_i^{n_i t} \\
&= \sum_{t \geq 1} \frac{1}{t} \left(\prod_{i=1}^{d} \sum_{n_i \geq 0} x_i^{n_i t} - 1\right)
\end{aligned}$$

For $t = 1$,

$$\prod_{i=1}^{d} \frac{1}{(1 - x_i^t)} - 1 < \frac{1}{\prod_{i=1}^{d} (1 - x_i)}.$$

For $t \geq 2$,

$$\begin{aligned}
\prod_{i=1}^{d} \frac{1}{(1 - x_i^t)} - 1 &= \frac{1 - \prod_{i=1}^{d} (1 - x_i^t)}{\prod_{i=1}^{d} (1 - x_i) \prod_{i=1}^{d} \sum_{j=0}^{t-1} x_i^j} \\
&\leq \frac{1}{\prod_{i=1}^{d} (1 - x_i)} \cdot \frac{\sum_{i=1}^{d} x_i^t}{\sum_{i=1}^{d} \sum_{j=1}^{t-1} x_i^j} \\
&\leq \frac{1}{\prod_{i=1}^{d} (1 - x_i)} \cdot \frac{\sum_{i=1}^{d} x_i^t}{\sum_{i=1}^{d} (t-1) x_i^t} \\
&= \frac{1}{\prod_{i=1}^{d} (1 - x_i)} \cdot \frac{1}{t-1}.
\end{aligned}$$

It follows that

$$\ln f(\bar{x}) \leq \frac{1 + \sum_{t \geq 2} \frac{1}{t(t-1)}}{\prod_{i=1}^{d} (1 - x_i)} = \frac{2}{\prod_{i=1}^{d} (1 - x_i)}. \tag{5.4}$$

Note that $p(\bar{n})$ is monotonically increasing. For a given $\bar{n}$,

$$f(\bar{x}) \geq \sum_{\bar{v} \geq \bar{n}} p(\bar{v}) \prod_{i=1}^{d} x_i^{v_i} \geq p(\bar{n}) \sum_{\bar{u} \geq 0} \prod_{i=1}^{d} x_i^{n_i+u_i} = p(\bar{n}) \prod_{i=1}^{d} \frac{x_i^{n_i}}{1-x_i}.$$

Hence

$$\ln p(\bar{n}) \leq \ln f(\bar{x}) + \sum_{i=1}^{d} (\ln(1-x_i) - n_i \ln x_i) \leq \ln f(\bar{x}) - \sum_{i=1}^{d} n_i \ln x_i.$$

Combining with Equation (5.4), we get

$$\ln p(\bar{n}) \leq \frac{2}{\prod_{i=1}^{d}(1-x_i)} - \sum_{i=1}^{d} n_i \ln x_i.$$

let $u_i = \frac{1}{1-x_i}$. Then

$$\ln p(\bar{n}) \leq g(\bar{u}) \stackrel{\text{def}}{=} 2 \prod_{i=1}^{d} u_i + \sum_{i=1}^{d} n_i \ln \frac{u_i}{u_i - 1}.$$

Taking partial derivatives, we have

$$\frac{\partial g(\bar{u})}{\partial u_i} = 2 \prod_{j \neq i} u_j - \frac{n_i}{u_i(u_i-1)},$$

Equating the partial derivatives to 0, we get

$$u_i = 1 + \frac{n_i}{2\boldsymbol{\pi}},$$

where $\boldsymbol{\pi} = \prod_{i=1}^{d} u_i$ can be found by solving equation

$$\boldsymbol{\pi} = \prod_{i=1}^{d} \left(1 + \frac{n_i}{2\boldsymbol{\pi}}\right).$$

Then

$$\ln p(\bar{n}) \leq 2\boldsymbol{\pi} + \sum_{i=1}^{d} n_i \log \left(1 + \frac{2\boldsymbol{\pi}}{n_i}\right).$$

We show that, for $n \geq 4$,

$$\boldsymbol{\pi} \leq n^{\frac{d}{d+1}}.$$

Let $n = \sum_{i=1}^{d} n_i$. By the concavity of the function $x \ln\left(1 + \frac{2\boldsymbol{\pi}}{x}\right)$, we have

$$\ln p(\bar{n}) \leq 2\boldsymbol{\pi} + n \ln \left(1 + \frac{2\boldsymbol{\pi} d}{n}\right).$$

Suppose $\boldsymbol{\pi} > n^{\frac{d}{d+1}}$. Then

$$n^{\frac{d}{d+1}} < \boldsymbol{\pi} \leq \left(1 + \frac{n}{2\boldsymbol{\pi}d}\right)^d \leq \left(1 + \frac{1}{2d} n^{\frac{1}{d+1}}\right)^d.$$

Solving for n, we get

$$n < \left(1 + \frac{1}{2d-1}\right)^{d+1} \leq 4,$$

a contradiction. Thus for $n \geq 4$ we must have $\boldsymbol{\pi} \leq n^{\frac{d}{d+1}}$. Hence

$$\ln p(\bar{n}) \leq 2n^{\frac{d}{d+1}} + n \ln\left(1 + 2d\, n^{-\frac{1}{d+1}}\right). \qquad \square$$

This complete the proof.

5.3 Even and Odd Graphs

Recall that, given a set of graphs $\mathcal{G}$, $\texttt{diff}(\mathcal{G})$ is the difference between the number of even and odd graphs. We prove Lemma 5.7, which states that

$$\texttt{diff}(\mathcal{C}_n) = (-1)^{n-1}(n-1)!,$$

where $\mathcal{C}_n$ is the set of connected graphs over the set $[n]$. We provide two recursive proofs and a third proof that reveals a relation between $\texttt{diff}(\mathcal{C}_n)$ and the enumeration of inversion-free trees.

5.3.1 Proof by Removing an Edge

We prove the following recursion and Lemma 5.7 follows as a corollary.

Proposition 5.8.

$$\texttt{diff}(\mathcal{C}_n) = -\sum_{\substack{n_1>0, n_2>0 \\ n_1+n_2=n}} \binom{n}{n_1} \texttt{diff}(\mathcal{C}_{n_1})\texttt{diff}(\mathcal{C}_{n_2}).$$

Proof. Let $e_0 = \{1, 2\}$. Then $\mathcal{C}_n$ can be partitioned into the following three sets:

- $\mathcal{C}_n^- \stackrel{\text{def}}{=} \{G \in \mathcal{C}_n \mid e_0 \notin G\}$.

- $\mathcal{C}_n^0 \stackrel{\text{def}}{=} \{G \in \mathcal{C}_n \mid e_0 \in G, G - e_0 \text{ is disconnected}\}$.
- $\mathcal{C}_n^+ \stackrel{\text{def}}{=} \{G \in \mathcal{C}_n \mid e_0 \in G, G - e_0 \text{ is connected}\}$.

Define a mapping $f : \mathcal{C}_n^- \to \mathcal{C}_n^+$ such that $f(G) = G + e_0$ for any $G \in \mathcal{C}_n^-$. it is easy to see that f is well-defined, and it's 1-1 and onto, where the inverse is $f^{-1}(G) = G - e_0$ for any $G \in \mathcal{C}_n^+$. Clearly $f(G)$ and G have different parities. It follows that f pairs up even graphs in $\mathcal{C}_n^-$ with odd graphs in $\mathcal{C}_n^+$, as well as odd graphs in $\mathcal{C}_n^-$ with even graphs in $\mathcal{C}_n^+$. Thus

$$\texttt{diff}(\mathcal{C}_n^-) + \texttt{diff}(\mathcal{C}_n^+) = \texttt{even}(\mathcal{C}_n^-) - \texttt{odd}(\mathcal{C}_n^-) + \texttt{even}(\mathcal{C}_n^+) - \texttt{odd}(\mathcal{C}_n^+) = 0,$$

and hence

$$\texttt{diff}(\mathcal{C}_n) = \texttt{diff}(\mathcal{C}_n^-) + \texttt{diff}(\mathcal{C}_n^0) + \texttt{diff}(\mathcal{C}_n^+) = \texttt{diff}(\mathcal{C}_n^0).$$

For any positive integers n_1 and n_2 such that $n = n_1 + n_2$, let

$$\mathcal{C}_{n_1,n_2} \stackrel{\text{def}}{=} \mathcal{C}_{[n_1]} \otimes \mathcal{C}_{[n_1+1..n]},$$

the set of graphs over $[n]$ with exactly two connected components whose vertex sets are $[n_1]$ and $[n_1 + 1 \,..\, n]$ respectively. Then

$$\texttt{diff}(\mathcal{C}_n^0) = -\sum_{\substack{n_1>0,n_2>0\\ n_1+n_2=n}} \binom{n}{n_1} \texttt{diff}(\mathcal{C}_{n_1,n_2}),$$

where the negative sign comes from e_0, and $\binom{n}{n_1}$ is the number of partitions of $[n]$ into two ordered subsets of sizes n_1 and n_2. The conclusion follows from that

$$\texttt{diff}(\mathcal{C}_{n_1,n_2}) = \texttt{diff}(\mathcal{C}_{n_1}) \cdot \texttt{diff}(\mathcal{C}_{n_2}). \qquad \square$$

Proof 1 of Lemma 5.7. By Proposition 5.8, it suffices to show that

$$(-1)^{n-1}(n-1)! = \sum_{\substack{n_1>0,n_2>0\\ n_1+n_2=n}} \binom{n}{n_1} (-1)^{n_1-1}(n_1-1)! \cdot (-1)^{n_2-1}(n_2-1)!,$$

which can be verified by straightforward calculations. $\square$

5.3.2 Proof by Removing a Vertex

Let Φ_n be the set of partitions of number n into positive integers; two partitions that differ only in the order of numbers are considered identical. For any partition $\varphi \in \Phi_n$, let φ_μ be the number of μ's in φ.

We prove the following recursion and Lemma 5.7 follows as a corollary.

Proposition 5.9.

$$\sum_{G \in \mathcal{C}_{n+1}} \mathtt{sign}\, G = \sum_{\varphi \in \Phi_n} \frac{n!}{\prod_\mu (\mu!)^{\varphi_\mu} \varphi_\mu!} \cdot \prod_\mu \left(- \sum_{G \in \mathcal{C}_\mu} \mathtt{sign}\, G \right)^{\varphi_\mu}.$$

Proof. Let $G \in \mathcal{C}_{n+1}$. Note that removing vertex $n+1$ from G results in a graph with one or more connected components. Thus G can be constructed as follows:

(1) Partition the set $[n]$ into non-empty subsets $V_1, V_2, \ldots, V_m$.

(2) Construct a connected graph G_i over each subset V_i.

(3) For each V_i, construct a claw-like tree T_i over $V_i \cup \{n+1\}$ with at least one edge such that all edges are incident to $n+1$.

It is easy to see that

$$\mathtt{sign}\, G = \prod_{i=1}^{m} \mathtt{sign}\, G_i \prod_{i=1}^{m} \mathtt{sign}\, T_i.$$

Given partition $= \{V_1, V_2, \ldots, V_m\}$ of $[n+1]$, let $\mathcal{C}(V_i)$ be the set of connected graphs on V_i and let $\mathcal{T}(V_i)$ be the set of trees over $V_i \cup \{n+1\}$ with at least one edge and all edges are incident to $n+1$. The steps of construction imply that the sum of $\mathtt{sign}\, G$ for all graphs constructed from the given partition is

$$\begin{aligned} \prod_{i=1}^{m} \left[\sum_{G \in \mathcal{C}(V_i)} \mathtt{sign}\, G \sum_{T \in \mathcal{T}(V_i)} \mathtt{sign}\, T \right] &= \prod_{i=1}^{m} \left[\sum_{G \in \mathcal{C}(V_i)} \mathtt{sign}\, G \cdot \sum_{e=1}^{|V_i|} \binom{|V_i|}{e} (-1)^e \right] \\ &= \prod_{i=1}^{m} \left(- \sum_{G \in \mathcal{C}_{|V_i|}} \mathtt{sign}\, G \right). \end{aligned}$$

Note that the right-hand side depends only on the sizes of V_i's. Let $\mu_i = |V_i|$. Then $\varphi = \{\mu_1, \mu_2, \ldots, \mu_m\}$ is a partition of the number n, and we can rewrite the right-hand side as

$$\prod_{\mu} \left(- \sum_{G \in \mathcal{C}_\mu} \mathtt{sign}\, G \right)^{\varphi_\mu}.$$

This quantity is for graphs constructed from a given partition of $[n]$. To complete the proof, note that the number of partitions of the set $[n]$ in which the sizes of the subsets form the same partition $\varphi = \{\mu_1, \mu_2, \ldots, \mu_m\}$ is

$$\frac{\binom{n}{\mu_1, \mu_2, \ldots, \mu_m}}{\prod_\mu \varphi_\mu!} = \frac{n!}{\prod_\mu (\mu!)^{\varphi_\mu} \varphi_\mu!},$$

and hence the conclusion follows. □

Proof 2 of Lemma 5.7. By Proposition 5.9, it suffices to show that

$$\sum_{\varphi \in \Phi_n} \frac{n!}{\prod_\mu (\mu!)^{\varphi_\mu} \varphi_\mu!} \cdot \prod_\mu (-1)^{\varphi_\mu} \left[(\mu - 1)!\right]^{\varphi_\mu} = (-1)^{n+1} n!,$$

or equivalently,

$$\sum_{\varphi \in \Phi_n} \frac{n!}{\prod_\mu (\mu!)^{\varphi_\mu} \varphi_\mu!} \cdot \prod_\mu \left[(\mu - 1)!\right]^{\varphi_\mu} = n!. \tag{5.5}$$

Note that the right-hand side $n!$ is the number of permutations of $[n]$. Since a permutation can be decomposed into cycles [Com74], it can be constructed by first partition $[n]$ into subsets then form a cycle from each subset. As mentioned in the proof of Proposition 5.9, there are $\frac{n!}{\prod_\mu (\mu!)^{\varphi_\mu} \varphi_\mu!}$ partitions of the set $[n]$ in which the sizes of the subsets form the same partition φ of the number n. Furthermore, there are $(\mu - 1)!$ ways to form a cycle from a subset of size μ. Thus the left-hand side of Equation (5.5) also counts the number of permutations. □

5.3.3 Proof by Inversion-free Trees

In a rooted tree over $[n]$, two adjacent vertices are called *parent* and *child*, where the parent has shorter distance to the root. Similarly, a pair of vertices are called *ancestor* and *descendant* if one of them is on the path from the root to the other vertex, where the ancestor is closer to the root.

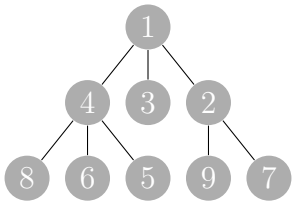


Figure 5.3: An Inversion-free Tree

Given a rooted tree over $[n]$, an *inversion* is a pair of ancestor and descendant vertices where the ancestor is larger. A rooted tree is *inversion-free* if every child is greater than its parent, as shown in Figure 5.3. Note that in inversion-frees every vertex is larger than all of its ancestors, and the root is always 1.

Let $f_n(i)$ be the number of rooted trees over $[n]$ with i inversions. Mallows and Riordan [MR68] determined the generating function of $f_n(i)$, showing in particular that

Proposition 5.10. *For all $n \geq 1$,*

$$f_n(0) = (n-1)!.$$

Here we provide a different, combinatorial, proof of this lemma.

Proof of Proposition 5.10. We define a 1-to-1 correspondence between inversion-free trees and permutations of the vertices whose first element is always the root.

Note that all the subtrees of an inversion-free tree are also inversion-free. Recursively define the mapping $\texttt{seq}(T_i)$ of the subtree T_i with root i and children $a_1 > a_2 > \cdots > a_\ell$ to be

$$\texttt{seq}(T_i) \stackrel{\text{def}}{=} i \cdot \texttt{seq}(T_{a_1}) \cdot \texttt{seq}(T_{a_2}) \cdot \cdots \cdot \texttt{seq}(T_{a_\ell})$$

For example, for the inversion-free tree T in Figure 5.3,

$$\texttt{seq}(T) = 1 \cdot \texttt{seq}(T_4) \cdot \texttt{seq}(T_3) \cdot \texttt{seq}(T_2) = 1\,4865\,3\,297.$$

Note that `seq` is essentially a *depth-first search (DFS)* traversal [KET06] on T, where children are visited in decreasing order.

It is easy to see that `seq` is 1-1. To show that it is onto, for any permutation $\bar{s} = s_0 s_1 \cdots s_{n'}$ of a subset $V \subseteq [n]$ where the first number s_0 is the smallest, we

construct an inversion-free tree $\texttt{tree}(\bar{s})$ as follows. Let $s[i..j]$ be the substring $s_i s_{i+1} \cdots s_j$. First decompose $\bar{s}[1..n']$ (excluding s_0) into substrings:

$$s[n_1 \,..\, n_2 - 1],\ s[n_2 \,..\, n_3 - 1],\ \ldots,\ s[n_t \,..\, n'],$$

where $n_1 = 1$, and $n_2, n_3, \ldots, n_t$ are recursively defined such that s_{n_i} is the first number in $s[n_{i-1} + 1 \,..\, n']$ that is smaller than $s[n_{i-1}]$. Clearly $s_{n_1} > s_{n_2} > \cdots > s_{n_t}$. The tree $\texttt{tree}(\bar{s})$ is recursively constructed to have s_0 as its root with $\texttt{tree}(s[n_1 \,..\, n_2 - 1])$, $\texttt{tree}(s[n_2 \,..\, n_3 - 1])$, $\ldots$, $\texttt{tree}(s[n_t \,..\, n'])$ attached to s_0.

For example, for the sequence $\bar{s} = 148653297$ where $s_0 = 1$ is the smallest number, $\bar{s}[1 \,..\, 8] = 48653297$ can be decomposed in the following steps: first we have $n_1 = 1$. The first number in $s[2 \,..\, 8] = 8653297$ less than $s_1 = 4$ is $s_5 = 3$. Therefore $n_2 = 5$ and our first substring is $s[1 \,..\, 4] = 4865$. Similarly, $n_3 = 6$. Thus we have the decomposition $s[1 \,..\, 8] = 4865 \cdot 3 \cdot 297$. Then $\texttt{tree}(\bar{s})$ can be constructed to have 1 as its root with $\texttt{tree}(4865), \texttt{tree}(3)$ and $\texttt{tree}(297)$ attached to 1, as shown in Figure 5.3.

It's easy to see that the two mappings we constructed are inverse of each other. Thus there is a one-to-one correspondence between the inversion-free trees and permutations of $[n]$ with 1 fixed. It follows that there are $(n-1)!$ inversion-free trees over $[n]$. □

Using Proposition 5.10, we show that $|C_n^e| - |C_n^o|$ differs from the number of inversion-free trees by at most a sign, which leads to another proof of Lemma 5.7.

Proof 3 of Lemma 5.7. Given a connected graph G over $[n]$, construct a rooted spanning tree $\texttt{span}(G)$ using the DFS starting from vertex 1, where neighbors are visited in decreasing order. Given a tree T over $[n]$, let $\mathcal{G}_T$ be the set of graphs G such that $\texttt{span}(G) = T$.

Given a tree T that is *not* inversion-free, define an automorphism on $\mathcal{G}_T$ as follows. First pick one inversion $\{u, v\}$ in T where u is an ancestor of v hence $u > v$. Note that by definition the root of T_G is always 1. Since $u > v$, $u \neq 1$.

Thus u is not the root and it has a parent, denoted as p_u. For any $G \in \mathcal{G}_T$, let

$$f(G) \stackrel{\text{def}}{=} \begin{cases} G - \{p_u, v\}, & \text{if } \{p_u, v\} \in G, \\ G + \{p_u, v\}, & \text{if } \{p_u, v\} \notin G. \end{cases}$$

By definition it's easy to see that $\texttt{span}(f(G)) = \texttt{span}(G) = T$, and that f is both 1-1 and onto. Clearly the numbers of edges in $f(G)$ and G have different parities. Thus $\texttt{even}(\mathcal{G}_T) = \texttt{odd}(\mathcal{G}_t)$, and hence $\texttt{diff}(\mathcal{G}_T) = 0$.

On the other hand, if T is inversion-free, we show that $\mathcal{G}_T = \{T\}$, i.e., the only graph whose spanning tree is T is itself. Suppose there exists $G \in \mathcal{G}_T$ such that $G \neq T$. Let $e = \{u, v\} \in G$ but $e \notin T$, where $u < v$. From the definition of DFS it is easy to see that there is no edge between two vertices that are not ancestor and descendant in $\texttt{span}(G)$. Thus u must be an ancestor of v in T. Let w be the child of u that is on the path to v. Then $u < w < v$. However, by of definition of our DFS v must be visited before w, namely v should also be a child of u, a contradiction. Therefore we have have $\mathcal{G}_T = \{T\}$. In conclusion,

$$\texttt{diff}(\mathcal{C}_n) = \texttt{diff}(\mathcal{T}_n^{\text{free}}) = (-1)^{n-1} \left|\mathcal{T}_n^{\text{free}}\right| = (-1)^{n-1}(n-1)!,$$

where $\mathcal{T}_n^{\text{free}}$ is the set of inversion-free trees over $[n]$, and we used Proposition 5.10 and the fact that any tree over n vertices has $n-1$ edges. □

Acknowledgment

Sections 5.1 and 5.2 appeared partially in Exact calculation of pattern probabilities, Jayadev Acharya, Hirakendu Das, Hosein Mohimani, Alon Orlitsky and Shengjun Pan, *IEEE International Symposium on Information Theory*, 2010. Section 5.3 will appear partially in On the number of even and odd connected graphs, Philippe Flajolet, Alon Orlitsky and Shengjun Pan, *In preparation*, 2012.

Chapter 6

Algorithms and Experiments

In Chapter 4 we showed how to calculate the PML distribution for pattern of special forms. For general pattern finding the exact PML distribution may be difficulty.

- In Section 6.1 we describe the procedure of an EM algorithm in [OSS+04, Zha05], and show that it is equivalent to a Generalized Gradient Ascend method.
- In Section 6.2 We will use the algorithm to evaluate the performance of PML on various distributions, and apply it to estimating the number of unseen symbols.

6.1 The Algorithms

It has been show [OSS+04, Zha05] that, given an initial support size k, the PML distribution can be approximated using the EM algorithm.

6.1.1 EM algorithm

Let $\mu_i(\bar{x})$ be the number of times the i-th symbol appears in the sequence $\bar{x}$. Starting with a arbitrary distribution $P = P^{(0)} = (p_1, p_2, \ldots, p_k)$, recursively update P as follows:

$$p_i \leftarrow \frac{1}{n}\mathbb{E}_{\bar{x}|\bar{\psi},P^{\text{old}}}\left[\mu_i(\bar{x})\right]. \tag{6.1}$$

For example, consider maximizing the probability of pattern 112 over all binary distributions $P = \{p_1, p_2\}$. Since $P(112) = p_1^2 p_2 + p_2^2 p_1 = p_1 p_2$, we have

$$\mathbb{E}_{\bar{x}|112,P}\left[\mu_1(\bar{x})\right] = \frac{p_1^2 p_2}{P(112)} \cdot 2 + \frac{p_2^2 p_1}{P(112)} = \frac{p_1^2 p_2}{p_1 p_2} + \frac{p_2^2 p_1}{p_1 p_2} = 1 + p_1,$$

and similarly $\mathbb{E}_{\bar{x}|112,P}\left[\mu_2(\bar{x})\right] = 1 + p_2$. Hence in the M-step P is updated as

$$p_1 = \frac{1 + p_1^{\text{old}}}{3}, \text{ and } p_2 = \frac{1 + p_2^{\text{old}}}{3}.$$

When the EM converges, we have

$$p_1 = \frac{1 + p_1}{3}, \text{ and } p_2 = \frac{1 + p_2}{3},$$

i.e., $p_1 = p_2 = \frac{1}{2}$.

6.1.2 EM v.s. Generalized Gradient Ascent

We show that the updating formula (6.1) can also be obtained by the generalized gradient ascent method. To maximize $\ln P(\bar{\psi})$ under the constraint $\sum_{i=1}^{k} p_i = 1$, consider maximizing the Lagrangian function

$$f(P, \lambda) \stackrel{\text{def}}{=} \ln P(\bar{\psi}) + \lambda \left(1 - \sum_{i=1}^{k} p_i\right).$$

The gradient of f is

$$\nabla f = \left(\frac{\partial f}{\partial p_i}, \frac{\partial f}{\partial p_i}, \ldots, \frac{\partial f}{\partial p_i}, 1 - \sum_{i=1}^{k} p_i\right),$$

where

$$\frac{\partial f}{\partial p_i} = \frac{1}{P(\bar{\psi})} \frac{\partial P(\bar{\psi})}{\partial p_i} - \lambda,$$

which can be written as

$$\frac{\partial f}{\partial p_i} = \frac{1}{P(\bar{\psi})} \cdot \frac{\partial P(\bar{\psi})}{\partial p_i} - \lambda = \frac{1}{P(\bar{\psi})} \sum_{t=1}^{m} \mu_t p_i^{\mu_t - 1} P(\bar{\psi}_t) - \lambda = \frac{1}{p_i} \mathbb{E}_{\bar{x}|\bar{\psi},P}[\mu_i(\bar{x})] - \lambda,$$

One form of the generalized gradient ascent updates P and λ iteratively as follows:

$$p_i = p_i^{\text{old}} + \gamma_i \left(\frac{1}{p_i} \mathbb{E}_{\bar{x}|\bar{\psi},P}[\mu_i(\bar{x})] - \lambda\right), \forall i \in [k],$$

$$\lambda = \lambda^{\text{old}} + \gamma_\lambda \left(1 - \sum_{i=1}^{k} p_i^{\text{old}}\right),$$

where $\gamma_i \geq 0$ and $\gamma_\lambda \geq 0$. If we initially choose p_i's such that $\sum_{i=1} p_i = 1$, $\lambda = n$, and at each iteration $\gamma_i = \frac{p_i}{n}$, then it is straightforward to verify that at the completion of any iteration $\sum_{i=1} p_i$ remains 1 and λ remains n. Thus p_i is updated to

$$p_i \leftarrow p_i^{\text{old}} + \frac{p_i^{\text{old}}}{n}\left(\frac{1}{p_i^{\text{old}}}\mathbb{E}_{\bar{x}|\bar{\psi},P^{\text{old}}}[\mu_i(\bar{x})] - n\right) = \frac{\mathbb{E}_{\bar{x}|\bar{\psi},P^{\text{old}}}[\mu_i(\bar{x})]}{n},$$

which is the same as the updating formula in Equation (6.1).

6.1.3 Metropolis Algorithm

The updating formula in Equation (6.1) can be written as

$$p_i \leftarrow \frac{\mathbb{E}_{\bar{x}|\bar{\psi},P^{\text{old}}}[\mu_i(\bar{x})]}{n} = \frac{1}{n}\sum_{\bar{x}\in\bar{\psi}} \frac{P^{\text{old}}(\bar{x})}{P^{\text{old}}(\bar{\psi})}\mu_i(\bar{x}).$$

A direct calculation of the summation is not a practical approach since in general there are exponentially many sequences $\bar{x}$ with the same pattern $\bar{\psi}$, and calculating $P^{\text{old}}(\bar{\psi})$ adds further difficulty [ADM+10].

Given canonical pattern $\bar{\psi}$ and probability multiset $P = \{p_1, p_2, \ldots, p_k\}$, We use the Markov chain Monte Carlo (MCMC) sampling method to estimate

$$\mathbb{E}_{\bar{x}|\bar{\psi},P}[\mu_i(\bar{x})] = \sum_{\bar{x}\in\bar{\psi}} \frac{P(\bar{x})}{P(\bar{\psi})}\mu_i(\bar{x}).$$

The idea is to create a Markov chain process such that the stationary distribution is $P(\bar{x} \mid \bar{\psi})$. The Metropolis algorithm constructs such random walks on graphs over sequences of the pattern $\bar{\psi}$.

More precisely, Define graph $G_{\bar{\psi}}$ as follows. The vertex set is the set of sequences having pattern $\bar{\psi}$. Two sequences $x_1^{\mu_1}x_2^{\mu_2}\cdots x_m^{\mu_m}$ and $y_1^{\mu_1}y_2^{\mu_2}\cdots y_m^{\mu_m}$ are adjacent if and only if

(a) they differ in exactly one symbol, *i.e.*$x_{i_1} \neq y_{i_1}$ for some $i_1 \in [m]$, and $x_\ell \neq y_\ell$ for all $\ell \neq i_1$,

(b) they have a pair of symbols swapped, *i.e.*$x_{i_1} = y_{i_2}$, $x_{i_2} = y_{i_1}$ for some $i_1 \neq i_2 \in [m]$ and $x_\ell = y_\ell$ for all $\ell \neq i_1, i_2$, or

(c) they are the same sequence (self-loops).

Define a random walk on $G_{\bar{\psi}}$ as described in Algorithm 1 [OSS$^+$04, Zha05], where at each step a neighbor $\bar{y} \neq \bar{x}$ of the current sequence $\bar{x}$ is chosen. The random walk proceeds to $\bar{y}$ if $P(\bar{y}) \geq P(\bar{x})$ or otherwise with probability $\frac{P(\bar{y})}{P(\bar{x})}$. It stays at $\bar{x}$ with the remaining probability.

Algorithm 1 Metropolis algorithm on $G_{\bar{\psi}}$

1. Start with a random sequence $\bar{x}_0 \in \bar{\psi}$.
2. **loop**
3. Let the current state be $\bar{x} = x_1^{\mu_1} x_2^{\mu_2} \cdots x_m^{\mu_m}$, and let $x_{m+1}, \ldots, x_k$ be the symbols not in $\bar{x}$.
4. Uniformly generate $i_1 \in [m]$.
   Uniformly generate $i_2 \in [k] \setminus \{i_1\}$.
5. **if** $i_2 > m$ **then**
6. Let $\bar{y}$ be $\bar{x}$ with x_{i_1} replaced by x_{i_2}.
7. **else**
8. Let $\bar{y}$ be $\bar{x}$ with x_{i_1} and x_{i_2} swapped.
9. **end if**
10. **if** $P(\bar{y})/P(\bar{x}) \geq 1$ **then**
11. Transit to $\bar{y}$.
12. **else**
13. Transit to $\bar{y}$ with probability $P(\bar{y})/P(\bar{x})$; otherwise stay at $\bar{x}$.
14. **end if**
15. **end loop**

Remarks:

(1) The weight associated with each sequence $\bar{x}$ is $w_{\bar{x}} = P(\bar{x})$.

(2) The selection probabilities are defined implicitly as follows. In Step 4, the random indices i_1 and i_2 are used to define the neighbor $\bar{y}$. If $i_2 \in [k]$ but $i_2 \notin [m]$, $\bar{x}$ and $\bar{y}$ differ in exactly one symbol. Then $\bar{y}$ is selected with

probability

$$\lambda_{\{\bar{x},\bar{y}\}} = \frac{1}{m} \cdot \frac{1}{k-1}.$$

If $j \in [m]$, $\bar{x}$ and $\bar{y}$ have two symbols swapped. Then $\bar{y}$ is selected with probability

$$\lambda_{\{\bar{x},\bar{y}\}} = \frac{1}{m} \cdot \frac{1}{k-1} \cdot 2.$$

(3) In Steps 10 and 13 of Algorithm 1, the ratio $\frac{P(\bar{y})}{P(\bar{x})}$ can be calculated as follows. If $\bar{y}$ is obtained from Step 6, i.e., by replacing all occurrences of x_{i_1} with x_{i_2}, then

$$\frac{P(\bar{y})}{P(\bar{x})} = \frac{p_{i_2}^{\mu_{i_1}}}{p_{i_1}^{\mu_{i_1}}} = \left(\frac{p_{i_2}}{p_{i_1}}\right)^{\mu_{i_1}}.$$

If $\bar{y}$ is obtained from Step 8, i.e., by swapping all occurrences of x_{i_1} and x_{i_2}, then

$$\frac{P(\bar{y})}{P(\bar{x})} = \frac{p_{i_2}^{\mu_{i_1}} p_{i_1}^{\mu_{i_2}}}{p_{i_1}^{\mu_{i_1}} p_{i_2}^{\mu_{i_2}}} = \left(\frac{p_{i_2}}{p_{i_1}}\right)^{\mu_{i_1}-\mu_{i_2}}.$$

Using Algorithm 1 to generate T sequences $\bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(T)}$, we can estimate $\mathbb{E}_{\bar{x}|\bar{\psi},P}[\mu_i(\bar{x})]$, Then

$$\mathbb{E}_{\bar{x}|\bar{\psi},P}[\mu_i(\bar{x})] \approx \frac{1}{T}\sum_{t=1}^{T} \mu_i\left(\bar{x}^{(t)}\right).$$

6.2 Experiments

We conduct the following experiments to demonstrate the performance of PML and show how it can be used to predict unseen symbols.

6.2.1 Probability Estimation

As we have seen in the introduction section, for a uniform distribution over 500 symbols, with only 1000 samples PML can approximate the probability multiset very closely. We show two more examples on different distributions.

Figure 6.1 shows a Zipf distribution, where the probabilities are C/i for $i = 50, 51, \ldots, 500$ and $C = \left(\sum_{i=50}^{500} \frac{1}{i}\right)^{-1}$. We sample 500 times with replacement. Note that compared to the support size the sample size is very small. As we

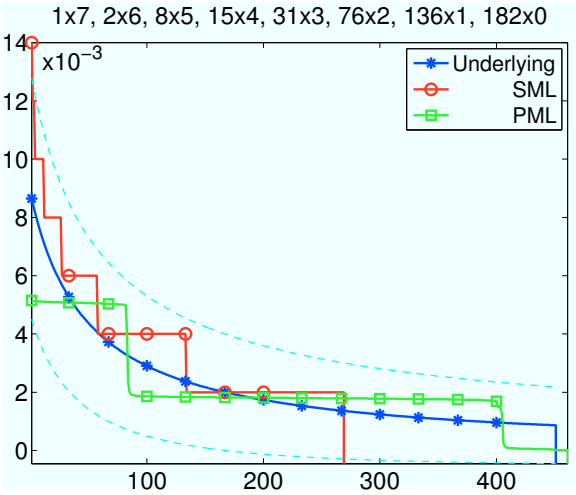


Figure 6.1: SML and PML Reconstructions for Zipf Distribution

Remark: The dashed lines are the Zipf probabilities off by one per-symbol standard deviation, i.e., $p_i \pm \sqrt{p_i(1-p_i)/n}$.

can see, SML overestimates large probabilities and it misses more than 2/5 of the symbols. On the other hand, PML approximates the probabilities fairly well and its estimate of the support size is very close to that of the underlying distribution.

Figure 6.2 shows the distribution over $k = 18,839$ last names from U.S. name census in 1999 consisting of 6,290,251 records. We sample $n = 35,000$ times from the distribution with replacement. The plot is in logarithm scale to show the subtle difference in small probabilities. Although SML performs well in estimating large probabilities, it misses nearly half of the names, while PML approximates more accurately not only the probabilities but also the support size.

6.2.2 Predicting New Symbols

In many applications the association between the probabilities and the underlying symbols is irrelevant. For example, the quantities such as support size, entropy, expected number of symbols of the same given frequency, etc., depend on only the probability multiset. PML can be potentially useful in such applications.

In 1985 a new poem, the *Taylor poem*, attributed to Shakespeare was discovered. To authenticate the authorship, Efron and Thisted [TE87] used a non-

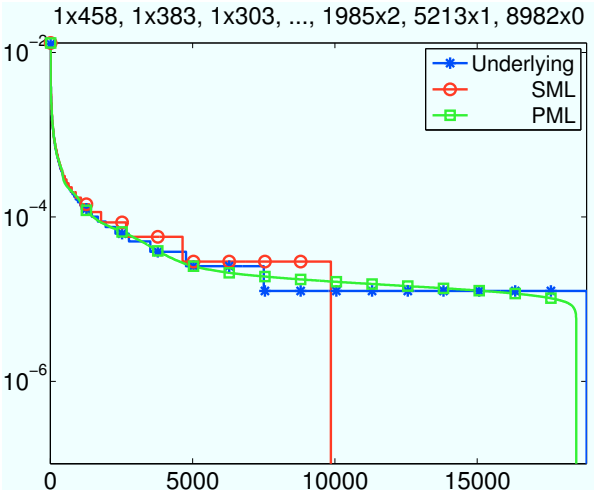


Figure 6.2: SML and PML Reconstructions for Name Distribution

parametric empirical Bayes model to examine the consistency of the word usage. For each $\mu \geq 0$, they estimate the expected number of distinct words in a new poem that appeared μ in Shakespeare's previous work.

More precisely, let $\bar{x}_1^n$ and $\bar{y}_1^N$ be two *i.i.d.* samples of sizes n and N respectively. Let m_μ be the number of distinct symbols in y_1^N that appear μ times in x_1^n. In the Taylor poem problem the two samples are Shakespeare's previous work and the new poem. Thisted and Efron estimated $\mathbb{E}[m_\mu]$, for $\mu \geq 0$, as follows

$$\tilde{m}_\mu \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} (-1)^{i+1} \binom{\mu+i}{i} \lambda^i \varphi_{\mu+i},$$

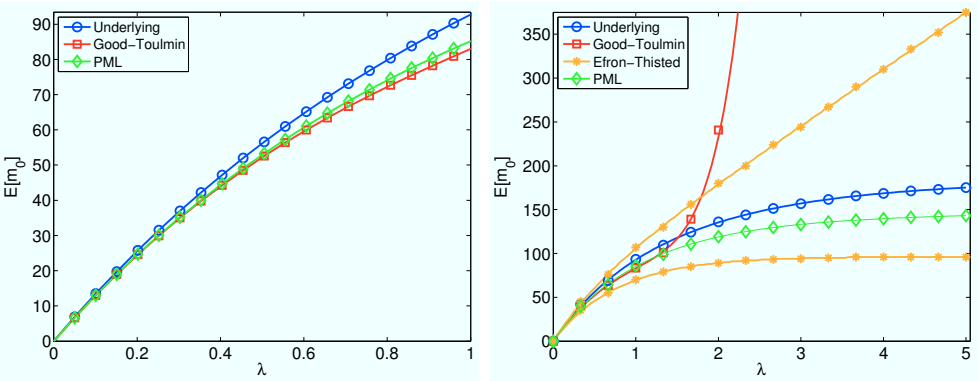
where $\lambda \stackrel{\text{def}}{=} N/n$ and $\varphi_{\mu+i}$ is the number of words that appear $\mu+i$ times in x_1^n. Then they compared $\tilde{m}_\mu$ to the actual value of m_μ.

Note that $\mathbb{E}[m_\mu]$ depends on only the probability multiset. If the probability multiset $P = \{p_1, p_2, \ldots, p_k\}$ of the underlying distribution is known, then $\mathbb{E}[m_\mu]$ can be calculated as

$$\mathbb{E}_P[m_\mu] = \sum_{i=1}^{k} \binom{n}{\mu} p_i^\mu (1-p_i)^{n-\mu} \big(1 - (1-p_i)^N\big).$$

Thus we can first use PML to obtain an estimate of the probabilities $\hat{P} = \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k\}$, then estimate $\mathbb{E}[m_\mu]$ as

$$\hat{m}_\mu \stackrel{\text{def}}{=} \mathbb{E}_{\hat{P}}[\mu_\mu].$$



(a) $\lambda < 1$ (b) $\lambda \geq 0$

Figure 6.3: Estimates of $\mathbb{E}[m_0]$ for Zipf Distribution

$\hat{m}_\mu$ on Zipf and Name Distributions

Since the true distribution of Shakespeare's vocabulary is unknown, to valuate the estimator $\hat{m}_\mu$, we first conduct experiments for $\hat{m}_0$ on the Zipf and name distributions from the previous subsection, and compare PML to the methods from Good-Toulmin [GT56] and Efron-Thisted [ET76, TE87].

For $\mu = 0$, Thisted and Efron's estimate becomes

$$\tilde{m}_0 = \lambda\varphi_1 - \lambda^2\varphi_2 + \lambda^3\varphi_3 - \lambda^4\varphi_4 + \cdots,$$

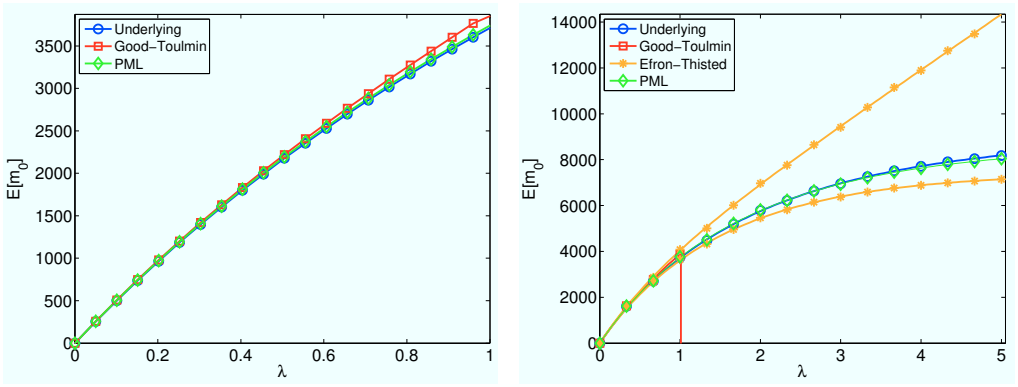
which is the same as Good-Toulmin [GT56].

Note that $\tilde{m}_0$ converges only if $\lambda < 1$. For $\lambda \geq 1$, Efron and Thisted [ET76] used linear programming to estimate a lower bound and an upper bound.

In each experiment, we take the sample $\bar{x}_1^n$ with replacement from the underlying distribution, and for various values of $\lambda = N/n$ we compare $\mathbb{E}[m_0]$, $\tilde{m}_0$, $\hat{m}_0$, and the lower and upper bounds from Efron and Thisted.

The results for the Zipf distribution are plotted in Figures 6.3(a) and 6.3(b). As we can see, for $\lambda < 1$, PML performs slightly better tahn the Good-Toulmin estimator $\tilde{m}_0$. For $\lambda \geq 1$ Good-Toulmin estimate diverges, while PML continues to perform well, and its prediction falls between the lower and upper bounds.

Similarly, for the name distribution PML outperforms $\hat{m}_0$ for both $\lambda < 1$ and $\lambda \geq 1$, as shown in Figures 6.4(a) and 6.4(b).



(a) $\lambda < 1$ (b) $\lambda \geq 0$

Figure 6.4: Estimates of $\mathbb{E}[m_0]$ for Name Distribution

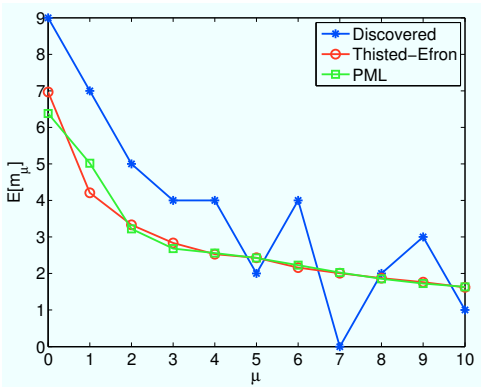


Figure 6.5: Estimates of $\mathbb{E}[m_\mu]$ for Shakespeare's Vocabulary

Back to Shakespeare

Shakespeare's previous work consists of $m = 31,534$ distinct words, and $n = 884,647$ in total, and the Taylor poem contains 258 distinct words and $N = 429$ in total [TE87].

We calculate the values of $\tilde{m}_\mu$ and $\hat{m}_\mu$ for μ up to 10, where $\lambda = N/n = 4.849 \times 10^{-4}$ is used in the calculation of $\tilde{m}_\mu$, and compare them to m_μ, the actual number of distinct words discovered in the Taylor poem that Shakespeare used μ times before. We we can see from the figure, PML is consistent with Efron and Thisted's method .

Acknowledgment

This chapter is adapted from Pattern maximum likelihood: computation and experiments, Alon Orlitsky, Shengjun Pan, Sajama, Narayana Prasad Santhanam, Krishnamurthy Viswanathan and Junan Zhang, *In preparation*, 2012.

Chapter 7

Set-Patterns

Recall that PML estimates the probability multiset of a single distribution. In this chapter we propose Set-pattern Maximum Likelihood (SPML) to estimate the probability multiset of concurrent random processes. we consider the scenario where we are given the samples of independent concurrent Bernoulli random processes.

- In Section 7.1 we define *set-pattern* and related notations.
- In Section 7.2 we show some basic properties of SPML.
- In Section 7.3 we show how to find the SPML for certain set-patterns.
- In Section 7.4 we develop an EM algorithm for general set-patterns.
- In Section 7.5 we consider a Poisson-version of set-pattern.

7.1 Notation and Definitions

Let $B_1, B_2, \ldots, B_k$ be independent Bernoulli distributions over alphabet $\{1, 0\}$, and the probabilities of 1 are $p_1, p_2, \ldots, p_k$. For each $i \in [k]$, independently sample T times from B_i. Denote all samples by a matrix $\boldsymbol{X}$ of size $T \times k$, where x_{ti} is the t-th sample from distribution B_i. The i-th *multiplicity* is

$$\mu_i(\boldsymbol{X}) \stackrel{\text{def}}{=} \sum_{t=1}^{T} x_{ti},$$

the number of times 1 is observed from distribution B_i. Denote the total number of ones in $\boldsymbol{X}$ by n. Then $n = \sum_{i=1}^{k} \mu_i(\boldsymbol{X})$.

It is easy to see that the probability of the sample matrix depends only on the multiplicities $\mu_i(\boldsymbol{X})$'s:

$$P(\boldsymbol{X}) = \prod_{i=1}^{k} \prod_{t=1}^{T} p_i^{x_{ti}} (1-p_i)^{1-x_{t_i}} = \prod_{i=1}^{k} p_i^{\mu_i(\boldsymbol{X})} (1-p_i)^{\mu_i(\boldsymbol{X})}.$$

For example, if $k = 3$, $T = 2$, and the samples are 11 from B_1, 11 from B_2, 01 from B_3 and 00 from B_4, then the sample matrix is

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix},$$

and its probability is

$$P(\boldsymbol{X}) = p_1^2 \cdot p_2^2 \cdot (1-p_3)^2 \cdot p_4(1-p_4).$$

Note that $\mu_i(\boldsymbol{X})$ could be 0. The *set-pattern* $\bar{\psi}(\boldsymbol{X})$ of $\boldsymbol{X}$ is the multiset of positive multiplicities, *i.e.*,

$$\bar{\psi}(\boldsymbol{X}) \stackrel{\text{def}}{=} \{\mu_i(\boldsymbol{X}) > 0 \mid i \in [k]\}^* .$$

Denote $\bar{\psi}(\boldsymbol{X})$ as

$$\bar{\psi}(\boldsymbol{X}) \stackrel{\text{def}}{=} 1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m},$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m > 0$. For example, the set-pattern of $\boldsymbol{X} = \left[\begin{smallmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{smallmatrix}\right]$, is $\bar{\psi}(\boldsymbol{X}) = 1^2 2^2 3$.

On the other hand, a set-pattern $\bar{\psi} = 1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m}$ can regarded as a set of all sample matrices whose set-pattern is $\bar{\psi}$. Slighting abusing the notations, we refer *multiplicities* of $\bar{\psi}$ to $\mu_1, \mu_2, \ldots, \mu_m$. For each $\mu \geq 0$, the *prevalence* φ_μ of μ is the number of instances μ in $\bar{\psi}$.

Given Bernoulli distributions with one's probabilities $p_1, p_2, \ldots, p_k$ and an integer $T > 0$, the *probability* of a set pattern $\bar{\psi}$ is the total probability of all sample matrices of size $T \times k$ that have set-pattern $\bar{\psi}$, *i.e.*,

$$P(\bar{\psi}, T) = \sum_{\boldsymbol{X}_{T\times K} : \bar{\psi}(\boldsymbol{X}) = \bar{\psi}} P(\boldsymbol{X}),$$

which can be rewritten as

$$P(\bar{\psi}, T) = \frac{1}{\prod_{\mu>0} \varphi_\mu!} \sum_{(i_1,i_2,\ldots,i_m)\in[k]^{\underline{m}}} \prod_{t=1}^{m} \binom{T}{\mu_t} p_{i_t}^{\mu_t} (1-p_{i_t})^{T-\mu_t} \cdot \prod_{j\notin\{i_1,i_2,\ldots,i_m\}} (1-p_j)^T,$$

where the division of $\prod_{\mu>0} \varphi_\mu!$ is to discount the overlap of mapping Bernoulli processes to multiplicities of the same value.

For example, given Bernoulli distributions with one's probabilities $p_1 = p_2 = p_3 = p_4 = p$, set-pattern $\bar{\psi} = 1^2 2^2 3$, and $T = 2$, the set-pattern probability is

$$\begin{aligned} P(1^2 2^2 3, 2) \cdot (1!2!) = &\binom{2}{2} p_1^2 \binom{2}{2} p_2^2 \binom{2}{1} p_3 (1-p_3) \cdot (1-p_4)^2 \\ &+ \binom{2}{2} p_1^2 \binom{2}{2} p_2^2 \binom{2}{1} p_4 (1-p_4) \cdot (1-p_3)^2 \\ &+ \cdots \\ &+ \binom{2}{2} p_4^2 \binom{2}{2} p_3^2 \binom{2}{1} p_2 (1-p_2) \cdot (1-p_1)^2, \end{aligned}$$

i.e.,

$$P(1^2 2^2 3, 2) = \frac{1}{2} \cdot 4^{\underline{3}} \cdot \binom{2}{2}^2 \binom{2}{1} p^5 (1-p)^3.$$

It's easy to see that for any set-pattern $\bar{\psi}$ and sampling time T, $P(\bar{\psi})$ is symmetric in p_i's, thus its value depends only on the multiset of multiplicities $\mathcal{M}(\bar{\psi})$ and the multiset of probabilities $P \stackrel{\text{def}}{=} \{p_1, p_2, \ldots, p_k\}^*$. Let

$$\mathscr{P}_d^{\text{sorted}} \stackrel{\text{def}}{=} \{(p_1, p_2, \ldots) \mid p_1 \geq p_2 \geq \cdots \geq 0\}.$$

Note $\mathcal{P}_d^{\text{sorted}} \subseteq \mathscr{P}_d^{\text{sorted}}$, where the latter does not require that $\sum_{i=1}^{\infty} p_i = 1$. Thus without loss of generality we may treat P as a vector in $\mathscr{P}_d^{\text{sorted}}$. Similar to PML, we define the *Set-pattern Maximum Likelihood* (SPML) of $\bar{\psi}$ and the correponding SPML multiset as

$$\hat{P}(\bar{\psi}, T) \stackrel{\text{def}}{=} \max_{P\in\mathscr{P}_d^{\text{sorted}}} P(\bar{\psi}, T), \text{ and } \hat{P}_{\bar{\psi},T} \stackrel{\text{def}}{=} \arg\max_{P\in\mathscr{P}_d^{\text{sorted}}} P(\bar{\psi}, T).$$

Again, we may run into the trouble that $\hat{P}_{\bar{\psi},T}$ does not exist. Instead of introducing a continuous part like we did for PML. In this dissertation we consider only discrete multisets. Namely we only study set-patterns for which the SPML multiset exists.

7.2 Properties

We first show how to rewrite the probability of a set-pattern in a simpler form, which will be used to prove the properties of SPML distributions, as well as the other results.

7.2.1 Reformulation

For any set-pattern $\bar{\psi}$ and $P = (p_1, p_2, \ldots, p_k) \in \mathscr{P}_d^{\text{sorted}}$,

$$\begin{aligned} P(\bar{\psi}, T) &= \frac{1}{\prod_{\mu \in \bar{\psi}} \varphi_\mu!} \sum_{\bar{i} \in [k]^{\underline{m}}} \prod_{t=1}^{m} \binom{T}{\mu_t} p_{i_t}^{\mu_t} (1 - p_{i_t})^{T - \mu_t} \cdot \prod_{i \in [k] \setminus \{i_1, i_2, \ldots, i_m\}} (1 - p_i)^T \\ &= \frac{1}{\prod_{\mu \in \bar{\psi}} \varphi_\mu!} \prod_{t=1}^{m} \binom{T}{\mu_t} \sum_{\bar{i} \in [k]^{\underline{m}}} \prod_{t=1}^{m} p_{i_t}^{\mu_t} (1 - p_{i_t})^{T - \mu_t} \cdot \prod_{i \in [k] \setminus \{i_1, i_2, \ldots, i_m\}} (1 - p_i)^T, \end{aligned}$$

If $0 < p_i < 1$ for all i, let

$$r_i \stackrel{\text{def}}{=} \frac{p_i}{1 - p_i}, \text{ and } R \stackrel{\text{def}}{=} (r_1, r_2, \ldots, r_k)\,.$$

The

$$P(\bar{\psi}, T) = \frac{1}{\prod_{\mu \in \bar{\psi}} \varphi_\mu!} \prod_{t=1}^{m} \binom{T}{\mu_t} \prod_{i=1}^{k} (1 - p_i)^T \sum_{\bar{i} \in [k]^{\underline{m}}} \prod_{t=1}^{m} \left(\frac{p_{i_t}}{1 - p_{i_t}} \right)^{\mu_t}.$$

Note that $1 - p_i = \frac{1}{1 + r_i}$. Then

$$P(\bar{\psi}, T) = \frac{1}{\prod_{\mu \in \bar{\psi}} \varphi_\mu!} \prod_{t=1}^{m} \binom{T}{\mu_t} \frac{\sum_{\bar{i} \in [k]^{\underline{m}}} r_{i_t}^{\mu_t}}{\prod_{i=1}^{k} (1 + r_i)^T} = \prod_{t=1}^{m} \binom{T}{\mu_t} \frac{R(\bar{\psi})}{\prod_{i=1}^{k} (1 + r_i)^T}, \tag{7.1}$$

where

$$R(\bar{\psi}) \stackrel{\text{def}}{=} \sum_{\bar{i} \in [k]^{\underline{m}}} \prod_{t=1}^{m} r_{i_t}^{\mu_t}.$$

Let

$$f(R, T) \stackrel{\text{def}}{=} \ln \frac{R(\bar{\psi})}{\prod_{i=1}^{k} (1 + r_i)^T} = \ln R(\bar{\psi}) - T \sum_{i=1}^{k} \ln(1 + r_i).$$

It follows that, for $\hat{P}_{\bar{\psi}, T}$ in which all probabilities are strictly less than 1,

$$\hat{P}_{\bar{\psi}, T} = \operatorname*{arg\,max}_{P \in \mathscr{P}_d^{\text{sorted}*}} f(R, T).$$

The reformulation of set-pattern probability $P(\bar{\psi})$ in Equation (7.1) requires that all probabilities in P are strictly less than 1. For a Bernoulli distribution with one's probability equal to 1, it is clear that an *i.i.d.* sample of size T has positive probability if and only if all observations are one. Thus in a set-pattern $\bar{\psi}$ with L multiplicities equal to T, its probability $P(\bar{\psi})$ is positive if and only P has at most L 1's.

Let $\ell \leq L$ be the number of 1's in P, *i.e.*,

$$p_1 = p_2 = \cdots = p_\ell = 1 > p_{\ell+1} \geq \cdots \geq p_k > 0.$$

Let

$$P^* \stackrel{\text{def}}{=} (p_\ell, p_{\ell+1}, \ldots, p_k).$$

Then

$$P(\bar{\psi}) = \varphi_T^\ell \, P^*(\bar{\psi}_{[\ell+1..m]}). \tag{7.2}$$

Let

$$\mathscr{P}_d^{\text{sorted}*} \stackrel{\text{def}}{=} \{(p_1, p_2, \ldots) \mid 1 > p_1 \geq p_2 \geq \cdots \geq 0\} \subseteq \mathscr{P}_d^{\text{sorted}}.$$

Thus the SPML distribution of a set-pattern with L multiplicities equal to T can be found by considering all distributions with up to L probabilities equal to one:

$$\hat{P}_{\bar{\psi},T} \stackrel{\text{def}}{=} \operatorname*{arg\,max}_{P=\mathbf{1}^\ell \cup P^* \colon P^* \in \mathscr{P}_d^{\text{sorted}*}} P^*(\bar{\psi}_{[\ell+1..m]}, T),$$

where $\mathbf{1}^\ell$ is a multiset of ℓ ones. Then for each ℓ the corresponding set-pattern probability $P^*(\bar{\psi}_{[\ell+1..m]}, T)$ can be reformulated using Equation (7.1).

7.2.2 Expansions

Similar to the expansion in Equation (2.4), we can write $R(\bar{\psi})$ as Recall that $R(\bar{\psi}) = \sum_{\bar{i} \in [k]^{\underline{m}}} \prod_{t=1}^m r_{i_t}^{\mu_t}$. Similar to the expansion in Equation (2.6), we can expand $R(\bar{\psi})$ as follows:

$$R(\bar{\psi}) = \sum_{S \subseteq [m]} R_I(\bar{\psi}_S) R_{\bar{I}}(\bar{\psi}_{\bar{S}}). \tag{7.3}$$

Particularly, for $I = \{i\}$,

$$R(\bar{\psi}) = R_i(\bar{\psi}) + \sum_{t=1}^m r_i^{\mu_t} R_i(\bar{\psi}_t). \tag{7.4}$$

7.2.3 Majorization

Given pattern $\bar{\psi}$, let

$$P_{\text{SML}} \stackrel{\text{def}}{=} \left(\frac{\mu_1}{T}, \frac{\mu_2}{T}, \ldots, \frac{\mu_m}{T}\right)$$

be the empirical distribution. Similar to the majorization property in Fact 2.3, we show that $\hat{P}_{\bar{\psi},T}$ is majorized by P_{SML}. Furthermore, we show that the total probability in $\hat{P}_{\bar{\psi},T}$ must be $\frac{n}{T}$.

Theorem 7.1. *For any set-pattern $\bar{\psi}$ such that $\hat{P}_{\bar{\psi},T}$ is discrete,*

(1) $\sum_{j=1}^{i} \hat{p}_j \le \frac{1}{T} \sum_{j=1}^{i} \mu_j, \forall i \le m$;

(2) $\sum_{j=1}^{k} \hat{p}_i = \frac{n}{T}$.

Proof. For simplicity let $P = \hat{P}_{\bar{\psi},T} = (p_1, p_2, \ldots, p_k)$. We first consider the case where $P \in \mathscr{P}_d^{\text{sorted}*}$, *i.e.*, $0 < p_i < 1$ for all $i \in [k]$. Recall that

$$r_i = \frac{p_i}{1-p_i}, R(\bar{\psi}) = \sum_{\bar{i} \in [k]^{\underline{m}}} \prod_{t=1}^{m} r_{i_t}^{\mu_t}, \text{ and } P(\bar{\psi}) = \frac{1}{\prod_{\mu \in \bar{\psi}} \varphi_\mu!} \prod_{t=1}^{m} \binom{T}{\mu_t} \frac{R(\bar{\psi})}{\prod_{i=1}^{k} (1+r_i)^T}.$$

Recall that

$$\hat{P}_{\bar{\psi},T} = \arg\max_{P \in \mathscr{P}_d^{\text{sorted}*}} f(R,T),$$

where

$$f(R,T) = \ln R(\bar{\psi}) - T \sum_{i=1}^{k} \ln(1+r_i).$$

Then For any $I \subseteq [k]$, let

$$R_I \stackrel{\text{def}}{=} \{r_i : i \in I\}.$$

By Expansion (7.3),

$$R(\bar{\psi}) = \sum_{S \subseteq [m]} R_I(\bar{\psi}_S) R_{\bar{I}}(\bar{\psi}_{\bar{S}}).$$

Note that R is not an actual distribution since r_i's don't have to sum up to 1.

Let R^α be obtained from R by scaling the r_i by α for all $i \in I$. Then

$$f(R^\alpha, T) = \ln \sum_{S \subseteq [m]} \alpha^{\mu_S} R_I(\bar{\psi}_S) R_{\bar{i}}(\bar{\psi}_{\bar{S}}) - T \sum_{i \in I} \ln(1+\alpha r_i) - T \sum_{i \in \bar{I}} \ln(1+r_i),$$

where $\mu_S = \sum_{t \in S} \mu_t$. Taking the derivative with respect to α we obtain

$$\frac{\partial f(R^\alpha, T)}{\partial \alpha} = \frac{1}{R^\alpha(\bar{\psi})} \sum_{S \subseteq [m], S \neq \emptyset} \mu_S \alpha^{\mu_S - 1} R_I(\bar{\psi}_S) R_{\bar{I}}(\bar{\psi}_{\bar{S}}) - T \sum_{i \in I} \frac{r_i}{1 + \alpha r_i}.$$

For $P = \hat{P}_{\bar{\psi},T}$, we must have $\left.\frac{\partial f(R^\alpha),T}{\partial \alpha}\right|_{\alpha=1} = 0$, *i.e.*,

$$\frac{1}{R(\bar{\psi})} \sum_{S \subseteq [m], S \neq \emptyset} \mu_S R_I(\bar{\psi}_S) R_{\bar{I}}(\bar{\psi}_{\bar{S}}) - T \sum_{i \in I} p_i = 0. \tag{7.5}$$

Note that $R_I(\bar{\psi}_S) = 0$ for any $S \subseteq [m]$ such that $|S| > |I|$. Thus in the expansion (7.3) we only need to consider $S \subseteq [m]$ such that $|S| \leq |I|$. Then

$$\begin{aligned}
\sum_{i \in I} p_i &= \frac{1}{T} \cdot \frac{1}{R(\bar{\psi})} \sum_{S \subseteq [m]:\, 0 < |S| \leq |I|} \mu_S R_I(\bar{\psi}_S) R_{\bar{I}}(\bar{\psi}_{\bar{S}}) \\
&\leq \frac{1}{T} \cdot \frac{\sum_{S \subseteq [m]:\, 0 < |S| \leq |I|} \mu_S R_I(\bar{\psi}_S) R_{\bar{I}}(\bar{\psi}_{\bar{S}})}{\sum_{S \subseteq [m]:\, 0 < |S| \leq |I|} R_I(\bar{\psi}_S) R_{\bar{I}}(\bar{\psi}_{\bar{S}})} \\
&\leq \frac{1}{T} \max_{S \subseteq [m]:\, 0 < |S| \leq |I|} \mu_S \\
&\leq \frac{1}{T} \sum_{i=1}^{|I|} \mu_i.
\end{aligned}$$

This proves the first part (under the assumption that all probabilities in $\hat{P}_{\bar{\psi},T}$ are less than 1). To show the second part, Let $I = [k]$. Then $\bar{I} = \emptyset$ and hence $R_{\bar{I}}(\bar{\psi}_{\bar{S}})$ for any $S \neq [m]$. It follows that Equation (7.5) becomes

$$\frac{1}{R(\bar{\psi})} \cdot \mu_{[m]} R(\bar{\psi}_{[m]}) - T \sum_{i=1}^{k} p_i = 0,$$

i.e., $\sum_{i=1}^{k} p_i = \frac{n}{T}$.

To finalize the proof, we show that the same results hold even if $P = \hat{P}_{\bar{\psi},T}$ has probabilities equal to 1. Let ℓ be the number of 1's in P and $P^* = (p_{\ell+1}, p_{\ell+2}, \ldots, p_k)$. Then we must have

$$\mu_1 = \mu_2 = \cdots = \mu_\ell = T.$$

Furthermore,

$$P(\bar{\psi}) = \varphi_T^{\ell} P^*(\bar{\psi}_{[\ell+1..m]}),$$

and P^* must maximize the probability of the set-pattern $\bar{\psi}_{[\ell+1..m]}$. Thus our previous results apply to P^*:

$$\sum_{j=\ell+1}^{i} p_j \leq \frac{1}{T} \sum_{j=\ell+1}^{i} \mu_j, \forall i \in [\ell+1, m],$$
$$\sum_{j=\ell+1}^{k} = \frac{1}{T} \sum_{j=\ell+1}^{m} \mu_j.$$

The conclusion follows by adding ℓ instances of 1 and $\frac{T}{T}$ to the left-hand and right-hand sides respectively. □

7.3 Set-patterns with Uniform SPML

We consider set-patterns with multiplicities sufficiently close so that the SPML distribution is uniform.

Similar to patterns, in a *uniform* set-pattern all multiplicities are the same. In a *quasi-uniform* set-pattern either $\mu_1 = \cdots = \mu_m = T$, or

$$\frac{T}{T - \mu_1}(\mu_t - \mu_{t'})^2 \leq \mu_t + \mu_{t'} - 2, \forall \{t, t'\} \in \binom{[m]}{2}.$$

It's easy to see that a uniform set-pattern is also quasi-uniform. We show that the SPML distribution of quasi-uniform set-patterns, hence uniform set-patterns is uniform.

Theorem 7.2. *For any quasi-uniform set-pattern $\bar{\psi}$ and sample time T,*

$$\hat{P}_{\bar{\psi},T} = \left(\frac{n}{\hat{k}T}, \ldots, \frac{n}{\hat{k}T}\right)$$

for some $\hat{k} \leq \infty$, where $n = \sum_{t=1}^{m} \mu_t$.

Proof. For simplicity let $P = \hat{P}_{\bar{\psi},T} = (p_1, p_2, \ldots, p_k)$. If there are more than one optimal P, choose one with smallest support size k.

If $\mu_1 = \mu_2 = \cdots = \mu_m = T$, it's easy to see that $P(\bar{\psi}, T)$ is maximized only if $k = m$ and $p_1 = p_2 = \cdots = p_m = 1$. Thus we may assume that $\mu_1 < T$, and for

all $\{t, t'\} \in \binom{[m]}{2}$

$$\frac{T}{T-\mu_1}(\mu_t - \mu_{t'})^2 \leq \mu_t + \mu_{t'} - 2.$$

It's easy to see that if $p_i = 1$ for some $i \in [k]$ then $P(\bar{\psi}, T) = 0$. Then we must have $P \in \mathscr{P}_d^{\text{sorted}*}$, and thus we must have

$$P = \hat{P}_{\bar{\psi},T} = \underset{P \in \mathscr{P}_d^{\text{sorted}*}}{\arg\max}\ f(R, T),$$

where

$$f(R, T) = \ln R(\bar{\psi}) - T\sum_{i=1}^{k} \ln(1 + r_i).$$

Since R maximizes $f(R, T)$, we must have, for all $i \in [k]$,

$$\frac{\partial f}{\partial r_i} = \frac{1}{R(\bar{\psi})} \cdot \frac{\partial R(\bar{\psi})}{\partial r_i} - \frac{T}{1 + r_i} = 0,$$

i.e.,

$$(1 + r_i) \cdot \frac{\partial R(\bar{\psi})}{\partial r_i} = TR(\bar{\psi}). \tag{7.6}$$

By Expansion (7.4), we can write $R(\bar{\psi})$ as

$$R(\bar{\psi}) = R_i(\bar{\psi}) + \sum_{t=1}^{m} r_i^{\mu_t} R_i(\bar{\psi}_t).$$

Then

$$\frac{\partial R(\bar{\psi})}{\partial r_i} = \sum_{t=1}^{m} \mu_t r_i^{\mu_t - 1} R_i(\bar{\psi}_t).$$

Thus Equation (7.6) becomes

$$\sum_{t=1}^{m} \mu_t (1 + r_i) r_i^{\mu_t - 1} R_i(\bar{\psi}_t) = TR(\bar{\psi}). \tag{7.7}$$

Similarly, for any $j \in [k] \setminus \{i\}$, we can write $R_i(\bar{\psi}_t)$ as

$$R_i(\bar{\psi}_t) = R_{i,j}(\bar{\psi}_t) + \sum_{t' \in [m] \setminus \{t\}} r_j^{\mu_{t'}} R_{i,j}(\bar{\psi}_{t,t'}).$$

thus

$$\sum_{(t,t'):\, t \neq t'} \mu_t (1 + r_i) r_i^{\mu_t - 1} r_j^{\mu_{t'}} R_{i,j}(\bar{\psi}_{t,t'}) + \sum_{t=1}^{m} \mu_t \left(r_i^{\mu_t} + r_i^{\mu_t - 1} \right) R_{i,j}(\bar{\psi}_t) = TR(\bar{\psi}). \tag{7.8}$$

Similarly,

$$\sum_{(t,t'):\, t\neq t'} \mu_t(1+r_j)r_j^{\mu_t-1}r_i^{\mu_{t'}}R_{ij}(\bar{\psi}_{t,t'}) + \sum_{t=1}^{m}\mu_t\left(r_j^{\mu_t}+r_j^{\mu_t-1}\right)R_{ij}(\bar{\psi}_t) = TR(\bar{\psi}). \quad (7.9)$$

Suppose P is not uniform. Choose i, j such that $p_i > p_j$. Then $r_i > r_j$. Using $(7.8) - (7.9)$, we get

$$\sum_{(t,t'):\, t\neq t'} \mu_t \cdot \frac{(1+r_i)r_i^{\mu_t-1}r_j^{\mu_{t'}} - (1+r_j)r_j^{\mu_t-1}r_i^{\mu_{t'}}}{r_i - r_j} R_{i,j}(\bar{\psi}_{t,t'}) + \sum_{t=1}^{m}\mu_t\left(\frac{r_i^{\mu_t}-r_j^{\mu_t}}{r_i-r_j} + \frac{r_i^{\mu_t-1}-r_j^{\mu_t-1}}{r_i-r_j}\right)R_{i,j}(\bar{\psi}_t) = 0. \quad (7.10)$$

On the other hand, Let $I = \{i, j\}$. Then by Expansion (7.3),

$$R(\bar{\psi}) = \sum_{(t,t'):\, t\neq t'} r_i^{\mu_t}r_j^{\mu_{t'}}R_{i,j}(\bar{\psi}_{t,t'}) + \sum_{t=1}^{m}(r_i^{\mu_t}+r_j^{\mu_t})R_{i,j}(\bar{\psi}_t) + R_{i,j}(\bar{\psi}). \quad (7.11)$$

Let $\tilde{P}$ be obtained from P with the following replacement:

$$\tilde{p}_i \leftarrow p_i + p_j - p_ip_j, \text{ and } \tilde{p}_j \leftarrow 0.$$

Then r_i and r_j are replaced by

$$\tilde{r}_i \leftarrow r_ir_j + r_i + r_j, \text{ and } \tilde{r}_j = 0.$$

It follows from Equation (7.11) that

$$\tilde{R}(\bar{\psi}) = \sum_{t=1}^{m}(r_ir_j + r_i + r_j)^{\mu_t}R_{i,j}(\bar{\psi}_t) + R_{i,j}(\bar{\psi}). \quad (7.12)$$

Note that $(1+\tilde{r}_i)(1+\tilde{r}_j) = (1+r_i)(1+r_j)$. Then

$$f(R,T) - f(\tilde{R},T) = \ln R(\bar{\psi}) - \ln \tilde{R}(\bar{\psi}) > 0.$$

i.e.,

$$\sum_{t=1}^{m}\left[(r_ir_j + r_i + r_j)^{\mu_t} - (r_i^{\mu_t}+r_j^{\mu_t})\right]R_{ij}(\bar{\psi}_t) < \sum_{(t,t')} r_i^{\mu_t}r_j^{\mu_{t'}}R_{ij}(\bar{\psi}_{t,t'}). \quad (7.13)$$

It can be shown that

Claim 7.1. *For any $r_i > r_j > 0$ and $\mu_t \geq 1$,*

$$(r_i r_j + r_i + r_j)^{\mu_t} - (r_i^{\mu_t} + r_j^{\mu_t}) \geq \mu_t (r_i r_j) \left(\frac{r_i^{\mu_t} - r_j^{\mu_t}}{r_i - r_j} + \frac{r_i^{\mu_t - 1} - r_j^{\mu_t - 1}}{r_i - r_j} \right).$$

See Appendix A.2 for the proof of Claim 7.1. Combining Equations (7.10), (7.13) and Claim 7.1, we get

$$\sum_{\{t,t'\} \in \binom{[m]}{2}} \left[(\mu_t - 1) \frac{r_i^{\mu_t - 1} r_j^{\mu_{t'}} - r_i^{\mu_{t'}} r_j^{\mu_t - 1}}{r_i - r_j} + (\mu_{t'} - 1) \frac{r_i^{\mu_{t'} - 1} r_j^{\mu_t} - r_i^{\mu_t} r_j^{\mu_{t'} - 1}}{r_i - r_j} \right.$$
$$\left. + (\mu_t - \mu_{t'}) \frac{r_i^{\mu_t} r_j^{\mu_{t'}} - r_i^{\mu_{t'}} r_j^{\mu_t}}{r_i - r_j} \right] R_{ij}(\bar{\psi}_{t,t'}) > 0, \quad (7.14)$$

We can show that

Claim 7.2. *For any $r_i > r_j > 0$ and $T > \mu_1 \geq \mu_t \geq \mu_{t'} \geq 0$,*

$$(\mu_t - 1) \frac{r_i^{\mu_t - 1} r_j^{\mu_{t'}} - r_i^{\mu_{t'}} r_j^{\mu_t - 1}}{r_i - r_j}$$
$$+ (\mu_{t'} - 1) \frac{r_i^{\mu_{t'} - 1} r_j^{\mu_t} - r_i^{\mu_t} r_j^{\mu_{t'} - 1}}{r_i - r_j} + (\mu_t - \mu_{t'}) \frac{r_i^{\mu_t} r_j^{\mu_{t'}} - r_i^{\mu_{t'}} r_j^{\mu_t}}{r_i - r_j}$$
$$\leq \left[\frac{T}{T - \mu_1} (\mu_t - \mu_{t'})^2 - (\mu_t + \mu_{t'} - 2) \right] (r_i^{\mu_t} r_j^{\mu_{t'}} + r_i^{\mu_{t'}} r_j^{\mu_t}) (r_i r_j)^{-1}. \quad (7.15)$$

See Appendix A.2 for the proof of Claim 7.2. Thus Inequality (7.14) implies that

$$\sum_{\{t,t'\} \in \binom{[m]}{2}} \left[\frac{T}{T - \mu_1} (\mu_t - \mu_{t'})^2 - (\mu_t + \mu_{t'} - 2) \right]$$
$$\cdot (r_i^{\mu_t} r_j^{\mu_{t'}} + r_i^{\mu_{t'}} r_j^{\mu_t}) (r_i r_j)^{-1} R_{ij}(\bar{\psi}_{t,t'}) > 0. \quad (7.16)$$

However, by the definition of quasi-uniform patterns, the left-hand side is nonnegative, a contradiction. □

Similar to almost-uniform patterns, Theorem 7.2 can be further generalized. Given set-pattern $\bar{\psi}$ such that $\mu_1 < T$ and $\{t, t'\} \in \binom{[m]}{2}$, let

$$D_{t,t'} \stackrel{\text{def}}{=} (\mu_t - \mu_{t'})^2 \frac{T}{T - \mu_1} - (\mu_t + \mu_{t'} - 2).$$

An *almost-uniform set-pattern* is either a quasi-uniform set-pattern, or $\mu_1 < T$ and

$$\sum_{D_{t,t'}<0} |D_{t,t'}| \cdot (\mu_m - 1)^{\mu_1-\mu_m} \geq \sum_{D_{t,t'}>0} D_{t,t'} \cdot \left[(\mu_1 - 1) + \frac{\mu_1}{T-\mu_1}(\mu_1 - \mu_m)\right]^{\mu_1-\mu_m}.$$

Theorem 7.3. *Given sample time T, the SPML distribution of an almost-uniform set-pattern $\bar{\psi}$ is*

$$\hat{P}_{\bar{\psi},T} = \left(\frac{n}{\hat{k}T}, \ldots, \frac{n}{\hat{k}T}\right)$$

for some $\hat{k} \leq \infty$, where $n = \sum_{t=1}^{m} \mu_t$.

See Appendix A.3 for the complete proof of Theorem 7.3.

7.3.1 Set-patterns with $\mu_1 = T$

As mentioned before, it is easy to see that, If the SPML distribution $P = \hat{P}_{\bar{\psi},T}$ of a set-pattern has ℓ probabilities, then $\bar{\psi}$ has at least ℓ multiplicities T equal to T. However, we give an example where the converse is not true.

Recall that, given $P \in \mathscr{P}_d^{\text{sorted}*}$ with ℓ ones and set-pattern $\bar{\psi}$ with L multiplicities equal to T, where $\ell \leq T$, we have

$$P(\bar{\psi}) = \varphi_T^{\ell}\, P^*(\bar{\psi}_{[\ell+1..m]}),$$

where $P^* = (p_{\ell+1}, p_{\ell+2}, \ldots, p_k)$. Consider the following set-pattern

$$\bar{\psi} = 1^5 2^3 3^3 \cdots 16^3.$$

Let $T = 5$ and $P = \hat{P}_{\bar{\psi},5} = (p_1, p_2, \ldots, p_k)$, if If $p_1 = 1$, then

$$P(\bar{\psi}, 5) = P^*(\bar{\psi}^*, 5).$$

Notice that $\bar{\psi}^*$ is uniform. Thus it follows from Theorem 7.2 that P^* is uniform, *i.e.*,

$$p_2 = p_3 = \cdots = p_k = p.$$

By Theorem 7.1,

$$\sum_{i=2}^{k} p_i = (k-1)p = \frac{n}{T} = 9,$$

i.e., $p = \frac{9}{k-1}$. Then

$$P(\bar{\psi}, 5) = \binom{k-1}{15} \left[p^3(1-p)^2\right]^{15} = \binom{k-1}{15} \left(\frac{9}{k-1}\right)^{45} \left(1 - \frac{9}{k-1}\right)^{30},$$

which is maximized to 1.1984×10^{-22} at $k = 16$.

On the other hand, consider the uniform distribution P with $p_1 = p_2 = \cdots = p_{16} = \frac{50}{5\times 16} = \frac{5}{8}$:

$$P(\bar{\psi}, 5) = 16 \left(\frac{50}{16}\right)^5 \binom{15}{15} \left[\left(\frac{50}{16}\right)^3 \left(1 - \frac{50}{16}\right)^2\right]^{15} = 1.6560 \times 10^{-22},$$

which is greater than 1.1984×10^{-22}, a contradiction. Thus the SPML distribution of $\bar{\psi} = 1^5 2^3 3^3 \cdots 16^3$ does not have probability 1.

7.4 Algorithm

Similar to patterns, we develop an EM algorithm to approximate the SPML multiset of set-patterns.

7.4.1 EM Algorithm

Given a canonical set-pattern $\bar{\psi} = 1^{\mu_1} 2^{\mu_2} \cdots m^{\mu_m}$ and a sampling time T, we are interested in the multiset of one's probabilities $P = (p_1, p_2, \ldots, p_k) \in \mathscr{P}_d^{\text{sorted}*}$ that maximizes the set-pattern probability $P(\bar{\psi}, T)$.

In general, given observation o, the EM algorithm estimates the underlying parameter θ that by maximizes the *log-likelihood* $\ln L(o, h \mid \theta)$, where h is the *hidden* variable. Starting with an initial guess of θ, the EM algorithm iteratively updates θ to increase $\ln L(o, h \mid \theta)$. Each iteration consists of two steps. The first step is to calculate

$$\mathbb{E}_{h|o,\theta^{\text{old}}} \left[\ln L(o, h \mid \theta)\right],$$

the expected value of the log-likelihood of the complete data, given the current parameter, with respect to the hidden variable conditional on the observation o and the previous parameter. The second step is to find the current parameter

θ maximizing the above expectation. In the case that o is a function of h, *i.e.*, $\Pr(o \mid h; \theta) = 1$, the log-likelihood becomes $\ln L(h \mid \theta)$ and hence the expectation can be simplified as

$$\mathbb{E}_{h|o,\theta^{\text{old}}} \left[\ln L(h \mid \theta)\right].$$

To apply the EM algorithm to set-patterns, the parameters are $P = (p_1, p_2, \ldots, p_k)$, the observation is the set-pattern $\bar{\psi}$, and the hidden variable is a sample matrix $\boldsymbol{X}$ with the given set-pattern. Note that $P(\bar{\psi} \mid \boldsymbol{X}) = 1$ and the log-likelihood is

$$\ln L(\boldsymbol{X} \mid P) = \ln P(\boldsymbol{X}).$$

Let P^{old} be the multiset from the previous iteration of the EM algorithm. To update P^{old}, we first calculate

$$\mathbb{E}_{\boldsymbol{X}|\bar{\psi},P^{\text{old}}} \left[\ln L(\boldsymbol{X} \mid P)\right] = \mathbb{E}_{\boldsymbol{X}|\bar{\psi},P^{\text{old}}} \left[\ln P(\boldsymbol{X})\right].$$

Without loss of generality, assume that $0 < p_i < 1$ for all $i \in [k]$. Recall that $r_i = \frac{p_i}{1-p_i}$. Then

$$P(\boldsymbol{X}) = \prod_{i=1}^{k} p_i^{\mu_i(\boldsymbol{X})} (1-p_i)^{T-\mu_i(\boldsymbol{X})} = \prod_{i=1}^{k} \frac{r_i^{\mu_i(\boldsymbol{X})}}{(1+r_i)^T}.$$

The expectation of the log-likelihood can then be written as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}|\bar{\psi},P^{\text{old}}} \left[\ln P(\boldsymbol{X})\right] &= \sum_{\boldsymbol{X}_{T\times k} \in \bar{\psi}} P^{\text{old}}(\boldsymbol{X} \mid \bar{\psi}) \left[\sum_{i=1}^{k} \mu_i(\boldsymbol{X}) \ln r_i\right] - T \sum_{i=1}^{k} \ln(1+r_i) \\
&= \sum_{i=1}^{k} \ln r_i \sum_{\boldsymbol{X} \in \bar{\psi}} \mu_i(\boldsymbol{X}) P^{\text{old}}(\boldsymbol{X} \mid \bar{\psi}) - T \sum_{i=1}^{k} \ln(1+r_i).
\end{aligned}$$

To maximize the expectation, let all partial derivatives with rep sect to r_i's be to 0, *i.e.*,

$$\frac{1}{r_i} \mathbb{E}_{\boldsymbol{X}|\bar{\psi},P^{\text{old}}} \left[\mu_i(\boldsymbol{X})\right] - \frac{T}{1+r_i} = 0.$$

Then

$$p_i = \frac{r_i}{1+r_i} = \frac{1}{T} \mathbb{E}_{\boldsymbol{X}|\bar{\psi},P^{\text{old}}} \left[\mu_i(\boldsymbol{X})\right]. \tag{7.17}$$

Note that, when the EM algorithm converges, we must have

$$p_i = \frac{1}{T} \mathbb{E}_{\boldsymbol{X}|\bar{\psi},P} \left[\mu_i(\boldsymbol{X})\right].$$

Then

$$\sum_{i=1}^{k} p_i = \frac{1}{T} \sum_{i=1}^{k} \mathbb{E}_{\boldsymbol{X}|\bar{\psi},P} \left[\mu_i(\boldsymbol{X})\right] = \frac{n}{T},$$

which is the same as the second part in Theorem 7.1.

7.4.2 Metropolis Algorithm

As show in the previous subsection, the SPML multiset can be approximated using the EM algorithm using updating formula (7.17). A direct calculation of the expectation as a summation is not a practical approach since in general there are exponentially many sample matrics $\boldsymbol{X}$ with the same set-pattern $\bar{\psi}$.

For any sample matrix $\boldsymbol{X}$, let the *sequence* of multiplicities

$$\bar{x} = \bar{\mu}(\boldsymbol{X}) \stackrel{\text{def}}{=} (\mu_1(\boldsymbol{X}), \mu_2(\boldsymbol{X}), \ldots, \mu_k(\boldsymbol{X}))\,.$$

Note that a sequence does not have the information of T. The probability of a sequence $\bar{x}$ with sample time T is

$$P(\bar{x}, T) = \sum_{\boldsymbol{X}_{T\times k}:\, \bar{\mu}(\boldsymbol{X})=\bar{x}} P(\boldsymbol{X}) = \prod_{i=1}^{k} \binom{T}{\mu_i(\bar{x})} p_i^{\mu_i(\bar{x})} (1-p_i)^{T-\mu_i(\bar{x})},$$

where $\mu_i(\bar{x})$ for the i-th component of $\bar{x}$. Then given canonical set-pattern $\bar{\psi}$ and probability multiset $P = (p_1, p_2, \ldots, p_k) \in \mathscr{P}_d^{\text{sorted}*}$, the expectation in the updating formula can be written as

$$\mathbb{E}_{\boldsymbol{X}|\bar{\psi},P} \left[\mu_i(\boldsymbol{X})\right] = \sum_{\boldsymbol{X}\in\bar{\psi}} \frac{P(\boldsymbol{X})}{P(\bar{\psi})} \mu_i(\boldsymbol{X}) = \sum_{\bar{x}\in\bar{\psi}} \frac{P(\bar{x}, T)}{P(\bar{\psi}, T)} \mu_i(\bar{x}),$$

where

We use the Markov chain Monte Carlo (MCMC) sampling method to estimate this summation. The idea is to use the Metropolis algorithm to create a random walk on the graph over the sequences $\bar{x}$, such that the stationary distribution is $P(\bar{x}|\bar{\psi})$.

Random Walks on Graphs

A *walk* on an undirected graph is a sequence of vertices, each adjacent to the next. A random walk on a graph G is defined by a transition probability

$P(v \to v')$ from every vertex v to each of its neighbors v'. Namely, $P(v \to v') \geq 0$, and

$$\sum_{v' \in N(v)} P(v \to v') = 1,$$

where $N(v)$ is the set of neighbors of v in G.

A graph is *connected* if there is walk from any vertex to any other. A graph is *aperiodic* if the greatest common divisor of all cycle lengths is 1. It is easy to see that every graph containing even one self-loop is aperiodic. The well-known Fundamental Theorem of Markov Chains says that every random walk on a connected aperiodic graph converges to a unique stationary distribution p, which satisfies, for all $v \in V$,

$$\sum_{v' \in N(v)} p(v')P(v' \to v) = p(v).$$

Let $G = (V, E)$ be a connected undirected graph with a self-loop at each vertex, where every vertex v is associated with a weight $w_v \geq 0$. The Metropolis algorithm creates a random walk over G by associating with every edge $\{v, v'\}$ a *selection probability* $\lambda_{\{v,v'\}}$ such that, for every vertex v,

$$\sum_{v' \in N(v) \setminus \{v\}} \lambda_{\{v,v'\}} \leq 1, \text{ and } \lambda_{\{v,v\}} \stackrel{\text{def}}{=} 1 - \sum_{v' \in N(v)} \lambda_{\{v,v'\}}.$$

The random walk then proceeds from each vertex v as follows:

- Select a random neighbor $v' \in N(v)$ according to its selection probability $\lambda_{\{v,v'\}}$.
- If $w_{v'} \geq w_v$, move to v', while if $w_{v'} < w_v$, move to v' with probability $\frac{w_{v'}}{w_v}$ and with the remaining probability stay at v.

It is easy to see that the transition probability from any vertex v to a neighbor v' is then

$$P(v \to v') = \lambda_{\{v,v'\}} \min\left\{1, \frac{w_{v'}}{w_v}\right\},$$

and

$$P(v \to v) \stackrel{\text{def}}{=} 1 - \sum_{v' \in N(v)} P(v \to v').$$

Note that G is connected, and since it contains self loops, it is also aperiodic. Hence the Fundamental Theorem of Markov Chains implies that our random walk convergences to a unique stationary distribution π. We show that

$$\pi(v) = \frac{w_v}{\sum_{u \in V} w_u}$$

as follows. Since π exists and is unique, it suffices to verify that

$$\sum_{v' \in N(v)} \frac{w_{v'}}{\sum_{u \in V} w_u} P(v' \to v) = \frac{w_v}{\sum_{u \in V} w_u},$$

or equivalently,

$$\sum_{v' \in N(v)} w_{v'} P(v' \to v) = w_v,$$

which is true since

$$\begin{aligned}
&\sum_{v' \in N(v)} w_{v'} P(v' \to v) \\
&= \sum_{v' \in N(v) \setminus \{v\}} w_{v'} P(v' \to v) + w_v P(v \to v) \\
&= \sum_{v' \in N(v) \setminus \{v\}} w_{v'} \lambda_{\{v',v\}} \min\left\{1, \frac{w_v}{w_{v'}}\right\} + w_v \left[1 - \sum_{v' \in N(v) \setminus \{v\}} \lambda_{\{v,v'\}} \min\left\{1, \frac{w_{v'}}{w_v}\right\}\right] \\
&= \sum_{v' \in N(v) \setminus \{v\}} \lambda_{\{v,'v\}} \min\{w_{v'}, w_v\} + w_v - \sum_{v' \in N(v) \setminus \{v\}} \lambda_{\{v,v'\}} \min\{w_v, w_{v'}\} \\
&= w_v.
\end{aligned}$$

Note that the algorithm does not require the calculation of $\sum_v w_v$, a prohibitive calculation when the graph is large. Furthermore, when every node of G has the same degree d, a natural selection probability is $\lambda_{\{v,v'\}} = 1/d$ for $v' \in N(v) \setminus \{v\}$. It is easily determined and ensures that a new node is always selected.

Estimating $\mathbb{E}_{\bar{\mu}|\bar{\psi},P}[\bar{\mu}_i]$

Define graph $G_{\bar{\psi}}$ as follows. The vertex set is the set of sequences having pattern $\bar{\psi}$. Two sequences $x_1^{\mu_1} x_2^{\mu_2} \cdots x_m^{\mu_m}$ and $y_1^{\mu_1} y_2^{\mu_2} \cdots y_m^{\mu_m}$ are adjacent if and only if

(a) they differ in exactly one symbol, *i.e.*$x_{i_1} \neq y_{i_1}$ for some $i_1 \in [m]$, and $x_\ell \neq y_\ell$ for all $\ell \neq i_1$,

(b) they have a pair of symbols swapped, *i.e.*$x_{i_1} = y_{i_2}$, $x_{i_2} = y_{i_1}$ for some $i_1 \neq i_2 \in [m]$ and $x_\ell = y_\ell$ for all $\ell \neq i_1, i_2$, or

(c) they are the same sequence (self-loops).

Define a random walk on $G_{\bar{\psi}}$ as described in Algorithm 2, where at each step a neighbor $\bar{y} \neq \bar{x}$ of the current sequence $\bar{x}$ is chosen. The random walk proceeds to $\bar{y}$ if $P(\bar{y}, T) \geq P(\bar{x}, T)$ or otherwise with probability $\frac{P(\bar{y})}{P(\bar{x},T)}$. It stays at $\bar{x}$ with the remaining probability.

Algorithm 2 Metropolis algorithm on $G_{\bar{\psi}}$

1. Start with a random sequence $\bar{x}_0 \in \bar{\psi}$.
2. **loop**
3. Let the current state be $\bar{x} = x_1^{\mu_1} x_2^{\mu_2} \cdots x_m^{\mu_m}$, and let $x_{m+1}, \ldots, x_k$ be the symbols not in $\bar{x}$.
4. Uniformly generate $i_1 \in [m]$.
   Uniformly generate $i_2 \in [k] \setminus \{i_1\}$.
5. **if** $i_2 > m$ **then**
6. Let $\bar{y}$ be $\bar{x}$ with x_{i_1} replaced by x_{i_2}.
7. **else**
8. Let $\bar{y}$ be $\bar{x}$ with x_{i_1} and x_{i_2} swapped.
9. **end if**
10. **if** $P(\bar{y}, T)/P(\bar{x}, T) \geq 1$ **then**
11. Transit to $\bar{y}$.
12. **else**
13. Transit to $\bar{y}$ with probability $P(\bar{y}, T)/P(\bar{x}, T)$; otherwise stay at $\bar{x}$.
14. **end if**
15. **end loop**

To estimate $\mathbb{E}_{\boldsymbol{X}|\bar{\psi},P}[\mu_i(\boldsymbol{X})]$, we use Algorithm 2 to generate N sequences $\bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(T)}$. Then

$$\mathbb{E}_{\boldsymbol{X}|\bar{\psi},P}[\mu_i(\boldsymbol{X})] = \sum_{\bar{x}\in\bar{\psi}} \frac{P(\bar{x},T)}{P(\bar{\psi})}\mu_i(\bar{x}) \approx \frac{1}{T}\sum_{t=1}^{T} \mu_i\big(\bar{x}^{(t)}\big)\,.$$

Remark: In Steps 10 and 13 of Algorithm 2, the ratio $\frac{P(\bar{y},T)}{P(\bar{x},T)}$ can be calculated as follows:

$$\frac{P(\bar{y},T)}{P(\bar{x},T)} = \frac{\prod_{i=1}^{k}\binom{T}{\mu_i(\bar{y})}p_i^{\mu_i(\bar{y})}(1-p_i)^{T-\mu_i(\bar{y})}}{\prod_{i=1}^{k}\binom{T}{\mu_i(\bar{x})}p_i^{\mu_i(\bar{x})}(1-p_i)^{T-\mu_i(\bar{x})}}.$$

Note that since both $\bar{x}$ and $\bar{y}$ satisfies the set-pattern $\bar{\psi}$, we have

$$\prod_{i=1}^{k}\binom{T}{\mu_i(\bar{x})} = \prod_{i=1}^{k}\binom{T}{\mu_i(\bar{y})} = \prod_{t=1}^{m}\binom{T}{\mu_t}.$$

Then

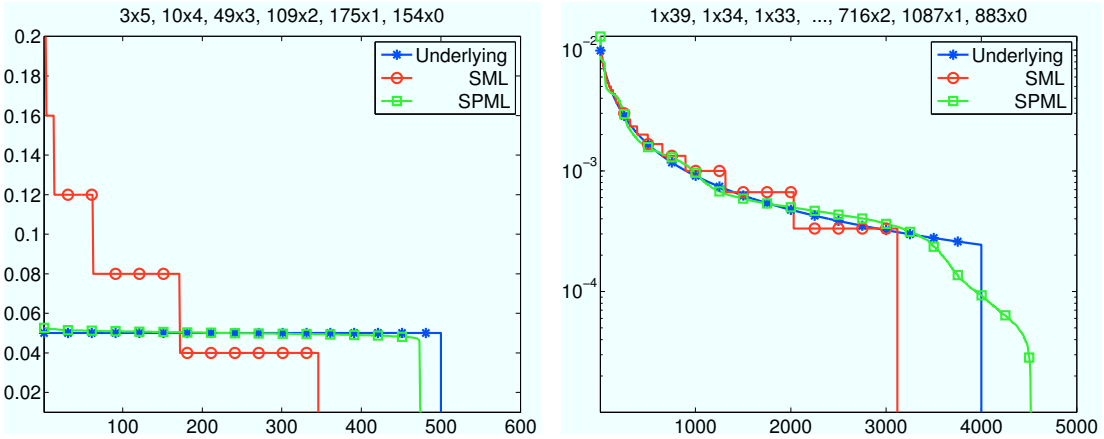
$$\frac{P(\bar{y},T)}{P(\bar{x},T)} = \frac{\prod_{i=1}^{k}p_i^{\mu_i(\bar{y})}(1-p_i)^{T-\mu_i(\bar{y})}}{\prod_{i=1}^{k}p_i^{\mu_i(\bar{x})}(1-p_i)^{T-\mu_i(\bar{x})}} = \frac{\prod_{i=1}^{k}r_i^{\mu_i(\bar{y})}}{\prod_{i=1}^{k}r_i^{\mu_i(\bar{x})}}.$$

- If $\bar{y}$ is obtained from Step 6, i.e., by replacing all occurrences of x_{i_1} with x_{i_2}, then
$$\frac{P(\bar{y})}{P(\bar{x})} = \left(\frac{r_{i_2}}{r_{i_1}}\right)^{\mu_{i_1}}.$$
- If $\bar{y}$ is obtained from Step 8, i.e., by swapping all occurrences of x_{i_1} and x_{i_2}, then
$$\frac{P(\bar{y})}{P(\bar{x})} = \frac{r_{i_2}^{\mu_{i_1}}r_{i_1}^{\mu_{i_2}}}{r_{i_1}^{\mu_{i_1}}r_{i_2}^{\mu_{i_2}}} = \left(\frac{r_{i_2}}{r_{i_1}}\right)^{\mu_{i_1}-\mu_{i_2}}.$$

7.4.3 Experiments

In the first experiment we take $k = 500$ identical distributions each with $p = 0.05$. We sample the set for $T = 25$ times.

From Figure 7.1(a) note that the SPML multiset is not only able to predict almost identical underlying multisets but also the values of p as well as the number of them. In comparison the empirical estimate is not only unable to predict the collection of Bernoulli distributions but it also misses 154 elements as it observes only 346 elements.



(a) $p_i = 0.05$ for $i = 1, 2, \ldots, 500$, $T = 25$ (b) $p_i = \frac{1}{100+i}$ for $i = 1, 2, \ldots, 4000, T = 3000$

Figure 7.1: Comparison between SML and SPML

The next experiment we run is when $p_i = \frac{1}{C+i}$, for $i = 1$ to k. This collection of values when normalized correspond to a zipf distribution. In the current set-up we take $C = 100$ and $k = 4000$. We sample from this collection of distributions $T = 3000$ times and from this sample we estimate the collection of underlying distributions.

From Figure 7.1(b) we note that the shape of set pml is much closer to the underlying distribution than the empirical distribution. One criterion we looked in order to compare is the ℓ_1 difference between the predicted and underlying distributions. In the case of empirical distribution the ℓ_1 distance is 0.6036 and for the SPML multiset it is 0.35.

7.5 Set-pattern with Poisson Processes

In the previous sections the concurrent processes are Bernoulli. If we consider Poisson processes with unit-time means $\lambda_1, \lambda_2, \ldots, \lambda_k$, since the time is continuous, we don't have a sample matrix. Instead, let the observation be the numbers of ones at time T:

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k).$$

The Poisson-version of the set-pattern can be defined similarly:

$$\bar{\psi}(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \{\mu \mid \mu > 0\}^* .$$

Then

$$\begin{aligned} P(\bar{\psi}, T) &= \frac{1}{\prod_{\mu>0} \varphi_\mu!} \sum_{(i_1,i_2,\ldots,i_m)\in[k]^{\underline{m}}} \prod_{t=1}^{m} e^{-\lambda_{i_t} T} \frac{(\lambda_{i_t} T)^{\mu_t}}{\mu_t!} \cdot \prod_{j\notin\{i_1,i_2,\ldots,i_m\}} e^{-\lambda_j T} \\ &= \frac{e^{-\sum_i \lambda_i T} T^n}{n!} \cdot \frac{\binom{n}{\mu_1,\mu_2,\ldots,\mu_m}}{\prod_{\mu>0} \varphi_\mu!} \sum_{(i_1,i_2,\ldots,i_m)\in[k]^{\underline{m}}} \prod_{i=1}^{m} \lambda_{i_t}^{\mu_t} \\ &= \frac{T^n}{n!} \cdot \frac{\binom{n}{\mu_1,\mu_2,\ldots,\mu_m}}{\prod_{\mu>0} \varphi_\mu!} \cdot \frac{(\sum_i \lambda_i)^n}{e^{\sum_i \lambda_i T}} P(\bar{\psi}), \end{aligned}$$

where

$$P(\bar{\psi}) = \sum_{(i_1,i_2,\ldots,i_m)\in[k]^{\underline{m}}} \prod_{t=1}^{m} \left(\frac{\lambda_{i_t}}{\sum_i \lambda_i}\right)^{\mu_t}$$

is the pattern probability under the distribution $P = \left(\frac{\lambda_1}{\sum_i \lambda_i}, \frac{\lambda_2}{\sum_i \lambda_i}, \ldots, \frac{\lambda_k}{\sum_i \lambda_i}\right)$. Note that $\frac{(\sum_i \lambda_i)^n}{e^{\sum_i \lambda_i T}}$ is maximized at $\sum_i \lambda_i = \frac{n}{T}$. It follows that maximizing the set-pattern probability with Poisson processes reduces to exactly PML.

Acknowledgment

This chapter appeared partially Estimating multiset of Bernoulli processes, Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky and Shengjun Pan, *Submitted to ISIT*, 2012.

Appendix A

Additional Proofs

This Chapter gathers technical proofs for completeness.

- Section A.1 lists the proofs of claims used for skewed patterns.
- Section A.2 lists the proofs of claims used for quasi-uniform set patterns.
- Section A.3 lists the proof for almost-uniform set-patterns.

A.1 Claims for Skewed Patterns

Claim 4.3 *For all* $r > 3$, $\hat{P}_{1^r2}(1^r2) \leq \frac{2}{5r}$.

Proof of Claim 4.3. As stated in Fact 2.12, the PML distribution of any binary pattern has support size 2. Let $\hat{P}_{1^r2} = (1-p, p)$, where $0 < p \leq \frac{1}{2}$. Then

$$\hat{P}P_{1^r2}(1^r2) = p^r(1-p) + (1-p)^r p.$$

(1) For $r = 4$, if $p \leq 0.32$, then

$$\hat{P}_{11112}(11112) \leq 0.32^4 \cdot (1-0.32) + (1-0.2)^4 \cdot 0.2 < \frac{2}{5 \cdot 4}.$$

If $0.32 < p \leq \frac{1}{2}$, then

$$\hat{P}_{11112}(11112) \leq 0.5^4 \cdot 0.5 + (1-0.32)^4 \cdot 0.32 < \frac{2}{5 \cdot 4}.$$

In either case, the claim holds.

(2) For $r \geq 5$, if $p \leq \frac{1}{3}$,

$$\begin{aligned}\hat{P}_{1^r2}(1^r2) &= p^r(1-p) + (1-p)^r p \\ &\leq \frac{1}{3^r} \cdot \frac{2}{3} + \left(\frac{r}{r+1}\right)^r \cdot \frac{1}{r+1} \\ &\leq \frac{1}{r}\left(\frac{2r}{3^{r+1}} + \frac{1}{e}\right) \\ &\leq \frac{1}{r}\left(\frac{2 \cdot 5}{3^{5+1}} + \frac{1}{e}\right) < \frac{2}{5r}.\end{aligned}$$

If $\frac{1}{3} < p \leq \frac{1}{2}$,

$$\begin{aligned}\hat{P}_{1^r2}(1^r2) &= p^r(1-p) + (1-p)^r p \\ &\leq 0.5^r \cdot 0.5 + \left(\frac{2}{3}\right)^r \cdot \frac{1}{3} \\ &= \frac{1}{r}\left(\frac{r}{2^{r+1}} + \frac{r2^r}{3^{r+1}}\right) \\ &\leq \frac{1}{r}\left(\frac{5}{2^{5+1}} + \frac{5 \cdot 2^5}{3^{5+1}}\right) < \frac{2}{5r}.\end{aligned}$$

Again, the claim holds in either case. □

Claim 4.4 *Given any $r \geq 3$ and $u \geq 2$, $F_{r,u}(p) > A_{r,u}$ for all $p \in (U_{r,u}, L_{r,u})$.*

Proof. Recall that

$$F_{r,u}(p) = \frac{r - (r+u)p - r(1-p)^{r+u}}{(r-1)u\,p(1-p)^{r+u-1}}.$$

Let

$$f(p) \stackrel{\text{def}}{=} A_{r,u}(r-1)u\,p(1-p)^{r+u-1} + (r+u)p + r(1-p)^{r+u} - r.$$

Then it's equivalent to show that $f(p) < 0$ for all $p \in (U_{r,u}, L_{r,u})$.

We first show that $f(U_{r,u}) < 0$ and $f(L_{r,u}) < 0$. Then we show that all $p \in (0,1)$ such that $f(p) < 0$ lies in a single interval and hence the conclusion follows.

(1) To show that $f(U_{r,u}) < 0$, we consider the following cases.

(a) For $r = 3$ and $u \in \{2, 3\}$ we can directly verify that $f(U_{r,u}) < 0$.

(b) If $r \geq 4$ and $u = 2$, then $U_{r,u} = \frac{r-1}{r+2}$ and $A_{r,u} = \left(\frac{r+u}{r}\right)^r \left(\frac{r+u}{u}\right)^u \cdot \frac{2}{5r}$. Hence

$$\begin{aligned}
f(H_{r,u}) &= f\left(\frac{r-1}{r+2}\right) \\
&= \left(\frac{r+2}{r}\right)^r \left(\frac{r+2}{2}\right)^2 \cdot \frac{2}{5r} \cdot 2(r-1) \cdot \frac{r-1}{r+2} \left(\frac{3}{r+2}\right)^{r+1} \\
&\quad + (r+2) \cdot \frac{r-1}{r+2} + r\left(\frac{3}{r+2}\right)^{r+2} - r \\
&= \frac{1}{5}(r-1)^2 \left(\frac{3}{r}\right)^{r+1} + r\left(\frac{3}{r+2}\right)^{r+2} - 1 \\
&\leq \frac{9}{5}\left(\frac{3}{r}\right)^{r-1} + 3\left(\frac{3}{r+2}\right)^{r+1} - 1 \\
&\leq \frac{9}{5}\left(\frac{3}{4}\right)^{4-1} + 3\left(\frac{3}{4+2}\right)^{4+1} - 1 \\
&< 0.
\end{aligned}$$

(c) If $r = 3$ and $u \geq 4$, or $r \geq 4$ and $u \geq 3$, then $U_{r,u} = \frac{r-1}{r+u}$ and $A_{r,u} = \left(\frac{r+u}{r}\right)^r \left(\frac{r+u}{u}\right)^u \left(\frac{r}{r+u-1}\right)^r \left(\frac{u}{r+u-1}\right)^{u-1}$. Hence

$$\begin{aligned}
f(U_{r,u}) &= f\left(\frac{r-1}{r+u}\right) \\
&= \left(\frac{r+u}{r}\right)^r \left(\frac{r+u}{u}\right)^u \cdot \left(\frac{r}{r+u-1}\right)^r \left(\frac{u-1}{r+u-1}\right)^{u-1} \\
&\quad \cdot \frac{(r-1)^2 u}{r+u} \left(\frac{u+1}{r+u}\right)^{r+u-1} + r\left(1 - \frac{r-1}{r+u}\right)^{r+u} + (r-1) - r \\
&= (r-1)^2 \left(\frac{u+1}{r+u-1}\right)^{r+u-1} \left(\frac{u-1}{u}\right)^{u-1} + r\left(\frac{u+1}{r+u}\right)^{r+u} - 1.
\end{aligned}$$

Let

$$g_1(r) \stackrel{\text{def}}{=} \ln\left[(r-1)^2 \left(\frac{u+1}{r+u-1}\right)^{r+u-1}\right],$$

$$g_2(r) \stackrel{\text{def}}{=} \ln\left[r\left(\frac{u+1}{r+u}\right)^{r+u}\right].$$

It is easy to see that

$$g_1'(r) = \ln\left(\frac{u+1}{r+u-1}\right) + \frac{2}{r-1} - 1 < 0,$$
$$g_2'(r) = \ln\left(\frac{u+1}{r+u}\right) + \frac{1}{r} - 1 < 0.$$

Then both $g_1(r)$ and $g_2(r)$ decrease in $r \geq 3$. It follows that

$$f(U_{r,u}) \leq 4\left(\frac{u+1}{u+2}\right)^{u+2}\left(\frac{u-1}{u}\right)^{u-1} + 3\left(\frac{u+1}{u+3}\right)^{u+3} - 1.$$

For $3 \leq u \leq 11$, we can directly verify that the right-hand side is negative. For $u \geq 12$,

$$f(U_{r,u}) \leq 4 \cdot \frac{1}{e} \cdot \frac{u}{e(u-1)} + \frac{3}{e^2} - 1 \leq \frac{4}{e^2} \cdot \frac{12}{11} + \frac{3}{e^2} - 1 < 0.$$

(2) To show that $f(L_{r,u}) < 0$, we consider the following cases.

(a) If $r = 3$ and $u \in \{2, 3\}$ we can directly verify that $f(L_{r,u}) < 0$.

(b) If $r \geq 4$ and $u = 2$, then $L_{r,u} = \frac{1}{r}$ and $A_{r,u} = \left(\frac{r+u}{r}\right)^r \left(\frac{r+u}{u}\right)^u \cdot \frac{2}{5r}$. Hence

$$f(L_{r,u}) = \left(\frac{r+2}{r}\right)^r \left(\frac{r+2}{2}\right)^2 \cdot \frac{2}{5r} \cdot 2(r-1) \cdot \frac{1}{r}\left(\frac{r-1}{r}\right)^{r+1} + \frac{r+2}{r} + r\left(\frac{r-1}{r}\right)^{r+2} - r.$$

For $4 \leq r \leq 33$, we can directly verify that the right-hand side is negative. for $r \geq 34$,

$$\begin{aligned} f(L_{r,u}) &\leq e^2\left(\frac{r+2}{2}\right)^2 \cdot \frac{2}{5r} \cdot \left(\frac{r-1}{r}\right)^{r+2} \\ &\quad + \frac{r+2}{r} + (r-1)\left(\frac{r-1}{r}\right)^{r+1} - r \\ &= \left[\frac{e^2(r+2)^2}{5r} + r - 1\right]\left(\frac{r-1}{r}\right)^{r+1} + \frac{r+2}{r} - r \\ &\leq \left[\frac{e^2(r+2)^2}{5r} + r - 1\right]e^{-1} + \frac{r+2}{r} - r \\ &= \frac{1}{5er}\left[-(5e - e^2 - 5)r^2 + (4e^2 + 5e - 5)r + (4e^2 + 10e)\right] \\ &< 0. \end{aligned}$$

(c) If $r = 3$ and $u \geq 4$, then $L_{r,u} = \frac{1}{4u}$ and

$$A_{r,u} = \left(\frac{r+u}{r}\right)^r \left(\frac{r+u}{u}\right)^u \left(\frac{r}{r+u-1}\right)^r \left(\frac{u}{r+u-1}\right)^{u-1}.$$

Hence

$$\begin{aligned} f(L_{r,u}) = &\left(\frac{u+3}{u+2}\right)^{u+2} \left(\frac{u-1}{u}\right)^{u-1} \cdot \frac{2(3+u)}{4u} \left(1 - \frac{1}{4u}\right)^{u+2} \\ &+ 3\left(1 - \frac{1}{4u}\right)^{u+3} + \frac{u+3}{4u} - 3. \end{aligned}$$

For $3 \leq u \leq 96$, we can directly verify that the right-hand side is negative. For $u \geq 97$,

$$\begin{aligned} f\left(\frac{1}{4u}\right) &\leq e\left(\frac{u}{u-1}\right) e^{-1} \cdot \frac{3+u}{2u} \left(1 - \frac{1}{4u}\right)^{u+2} \\ &\quad + 3\left(1 - \frac{1}{4u}\right)^{u+2} + \frac{3}{4u} - \frac{11}{4} \\ &= \frac{7u-3}{2(u-1)} \left(1 - \frac{1}{4u}\right)^{u+2} + \frac{3}{4u} - \frac{11}{4} \\ &\leq \left(\frac{7}{2} + \frac{2}{u-1}\right) \cdot e^{-\frac{u+2}{4u}} + \frac{3}{4u} - \frac{11}{4} \\ &\leq \left(\frac{7}{2} + \frac{2}{97-1}\right) \cdot e^{-\frac{1}{4}} + \frac{3}{4 \cdot 97} - \frac{11}{4} < 0. \end{aligned}$$

(d) If $r \geq 4$ and $u \geq 3$, then $L_{r,u} = \frac{1}{r+u}$ and

$$A_{r,u} = \left(\frac{r+u}{r}\right)^r \left(\frac{r+u}{u}\right)^u \left(\frac{r}{r+u-1}\right)^r \left(\frac{u}{r+u-1}\right)^{u-1}.$$

Hence

$$\begin{aligned} f(L_{r,u}) = &\left(\frac{r+u}{r}\right)^r \left(\frac{r+u}{u}\right)^u \cdot \left(\frac{r}{r+u-1}\right)^r \left(\frac{u-1}{r+u-1}\right)^{u-1} \\ &\cdot \frac{(r-1)u}{r+u} \left(\frac{r+u-1}{r+u}\right)^{r+u-1} + r\left(1 - \frac{1}{r+u}\right)^{r+u} + 1 - r \\ = &(r-1)\left(\frac{u-1}{u}\right)^{u-1} + r\left(1 - \frac{1}{r+u}\right)^{r+u} + 1 - r \\ \leq &(r-1)\left(\frac{3-1}{3}\right)^{3-1} + \frac{r}{e} + 1 - r \\ = &\frac{5e - (5e-9)r}{9} < 0. \end{aligned}$$

We have shown that $f(U_{r,u}) < 0$ and $f(L_{r,u}) < 0$. To complete the proof, we show that the value of $p \in (0,1)$ that satisfies $f(p) < 0$ lies in a single interval. Consider the derivatives of $f(p)$:

$$\begin{aligned} f'(p) &= (r+u) + (1-p)^{r+u} \\ &\qquad \cdot \left[(r+u)(r-(r-1)uA_{r,u})p + (r-1)uA_{r,u}\right], \\ f''(p) &= (r+u-1)(1-p)^{r+u-3} \\ &\qquad \cdot \left[(r+u)((r-1)uA_{r,u} - r)p + r(r+u) - 2(r-1)uA_{r,u}\right]. \end{aligned}$$

Note that, for any $0 < p < 1$, $f''(p) \geq 0$ if and only if

$$(r+u)((r-1)uA_{r,u} - r)p + r(r+u) - 2(r-1)uA_{r,u} \geq 0.$$

Furthermore, we can show that $(r-1)uA_{r,u} - r > 0$ as follows. Proposition 4.7 implies that

$$A_{r,u} \geq \frac{\hat{P}_{1^r23\cdots u}(1^r23\cdots u)}{\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1)} \geq \frac{\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u)}{\hat{P}_{1^r23\cdots u+1}(1^r23\cdots u+1)} \geq 1.$$

Thus

$$(r-1)uA_{r,u} - r \geq (r-1)u - r > 0.$$

Let

$$p^* \stackrel{\text{def}}{=} \frac{2(r-1)A_{r,u} - r(r+u)}{(r+u)[(r-1)uA_{r,u} - r]}.$$

Then, for $0 < p < 1$, $f''(p) \geq 0$ if and only if $p \geq p^*$. Note that

$$p^* = \frac{2}{(r+u)u} + \frac{2r/u - (r+u)}{(r+u)\left[(r-1)uA_{r,u} - r\right]} < \frac{2}{(r+u)u} < 1.$$

Suppose $p^* \leq 0$. Then $f'(p)$ is increasing for all $p \in (0,1)$. However, since

$$f'(0) = (r+u) + (r-1)uA_{r,u} > r + u = f'(1),$$

$f'(p)$ can't be monotonically increasing in $(0,1)$; we must have $0 < p^* < 1$. It follows that $f'(p)$ decreases in $(0,p^*)$, then increases in $(p^*,1)$. There are two possibilities. Let

(1) If $f'(p^*) \geq 0$, then $f'(p) \geq 0$ for all $p \in (0,1)$, and hence $f(p)$ in monotonically increasing in $p \in (0,1)$.

(2) If $f'(p^*) < 0$, then there exist $\alpha \in (0, p^*)$ and $\beta \in (p^*, 1)$ such that $f'(p) > 0$ for $p \in (0, \alpha) \cup (\beta, 1)$ and $f'(p) < 0$ for $p \in (\alpha, \beta)$. Then $f(p)$ first increases in $p \in (0, \alpha)$, then decreases in $p \in (\alpha, \beta)$, and increases again in $p \in (\beta, 1)$.

Note that $f(0) = 0$ and $f(1) = u > 0$. It follows that in either case $\{p \in (0, 1) \mid f(p) < 0\}$ is a single interval. We have previously shown that $f(U_{r,u}) < 0$ and $f(L_{r,u}) < 0$. Thus $f(p) < 0$ for any $p \in (U_{r,u}, L_{r,u})$. □

Claim 4.5 *For all $m \geq 2$, $\hat{P}_{1123\cdots m}(1123\cdots m) \leq \frac{2/e}{m(m-1)}$.*

Proof. As stated in Chapter 2, $\hat{P}_{1123\cdots m}$ is uniform. Let $k = \hat{k}$ be the support size. Then

$$\hat{P}_{1123\cdots m}(1123\cdots m) = k^{\underline{m}} \left(\frac{1}{k}\right)^{m+1} = \frac{1}{k} \cdot \prod_{i=1}^{m-1} \left(1 - \frac{i}{k}\right).$$

For any real number $x \in (0, 1)$, $1 - x < e^{-x}$. Thus

$$\hat{P}_{1123\cdots m}(1123\cdots m) \leq \frac{1}{k} \cdot \exp\left(-\frac{m(m-1)}{2k}\right) = f\left(\frac{1}{k}\right),$$

where $f(x) \stackrel{\text{def}}{=} x \exp\left(-\frac{m(m-1)}{2}x\right)$. it is easy to see that $f(x)$ is maximized at $x = \frac{2}{m(m-1)}$. Thus

$$\hat{P}_{1123\cdots m}(1123\cdots m) \leq f\left(\frac{2}{m(m-1)}\right) = \frac{2/e}{m(m-1)}.$$ □

Claim 4.6 *If $r \geq 3$ and $u \geq 2$, but $r \neq 3$ or $u \neq 2$, then*

$$\left(\frac{r}{r+u}\right)^r \left(\frac{u}{r+u}\right)^u \cdot \frac{u(u+1)}{2} > L_{r,u}^{r-2}.$$

Proof. It's equivalent to prove that

$$f(r, u) \stackrel{\text{def}}{=} \frac{1}{L_{r,u}^{r-2}} \left(\frac{r}{r+u}\right)^r \left(\frac{u}{r+u}\right)^u \frac{u(u+1)}{2} > 1.$$

(1) If $r = 3$ and $u = 3$,

$$f(r,u) = \frac{1}{0.07869}\left(\frac{3}{6}\right)^3\left(\frac{3}{6}\right)^3\frac{3\cdot 4}{2} > 1.$$

(2) If $r \geq 4$ and $u = 2$,

$$f(r,u) = r^{r-2}\cdot 3\left(\frac{r}{r+2}\right)\left(\frac{2}{r+2}\right)^2.$$

For $r = 4$, we can directly verify that the right-hand side is greater than 1. For $r \geq 5$,

$$f(r,u) \geq r^{r-4}\cdot\frac{3}{e^2}\left(\frac{2r}{r+2}\right)^2 \geq 5^{5-1}\cdot\frac{3}{e^2}\left(\frac{2\cdot 5}{5+2}\right)^2 > 1.$$

(3) If $r = 3$ and $u \geq 4$, $L_{r,u} = \frac{1}{4u}$. For $3 \leq u \leq 6$, we can directly verify that $f(3,u) > 1$. For $u \geq 7$,

$$f(3,u) = \frac{54}{e^3}\left(\frac{u}{u+3}\right)^2\frac{u+1}{u+3} > \frac{54\cdot 49\cdot(7+1)}{e^3(7+3)^3} > 1.$$

(4) If $r \geq 4$ and $u \geq 3$, $L_{r,u} = \frac{1}{r+u}$. Then

$$\begin{aligned} f(r,u) &= (r+u)^{r-2}\left(\frac{r}{r+u}\right)^r\left(\frac{u}{r+u}\right)^u\left(\frac{u(u+1)}{2}\right) \\ &= \frac{u^{u+1}(u+1)}{2}\cdot\frac{r^r}{(r+u)^{u+2}}. \end{aligned}$$

Let $g(r) \stackrel{\text{def}}{=} \ln\left[\frac{r^r}{(r+u)^{u+2}}\right]$. Since $g'(r) = \frac{r-2}{r+u} + \ln r > 0$, $g(r)$ increases. Hence

$$\begin{aligned} f(r,u) &\geq u^{u+1}(u+1)\cdot\frac{128}{(u+4)^{u+2}} \\ &= \left(\frac{u}{u+4}\right)^u\cdot\frac{128u(u+1)}{(u+4)^2}. \end{aligned}$$

For $3 \leq u \leq 6$, we can directly verify that the right-hand side is greater than 1. For $u \geq 7$,

$$f(r,u) \geq \frac{1}{e^4}\cdot\frac{128\cdot 7\cdot 8}{(7+4)^2} > 1. \qquad \square$$

Claim 4.7 *If $r \geq 3$ and $u \geq 2$, but $r \neq 3$ or $u \neq 2$, then*

$$f_{r,u}(p) \stackrel{\text{def}}{=} p^{r-2}\left[\frac{(r+1)u}{r+u} - p\right](1 - U_{r,u} - p)$$

increases for $p \in (0, L_{r,u}]$.

Proof. We rewrite $f_{r,u}(p)$ as the product two parts:

$$f_{r,u}(p) = g_{r,u}(p) \cdot h_{r,u}(p),$$

where

$$g_{r,u}(p) = p^{\frac{1}{2}}\left[\frac{(r+1)u}{r+u} - p\right], \text{ and } h_{r,u}(p) = p^{r-\frac{5}{2}}(1 - U_{r,u} - p).$$

It's easy to show that $g_{r,u}(p)$ increases for $p \leq \frac{1}{3}\frac{(r+1)u}{r+u}$, and $h_{r,u}(p)$ increases for $p \leq \frac{2r-5}{2r-3}(1 - U_{r,u})$. Thus it's sufficient to show that

$$L_{r,u} \leq \frac{1}{3}\frac{(r+1)u}{r+u}, \text{ and } L_{r,u} \leq \frac{2r-5}{2r-3}(1 - U_{r,u}).$$

(1) If $r = 3$ and $u = 3$, then $L_{r,u} = 0.07869$ and $U_{r,u} = 0.4199$. Hence

$$\frac{\frac{1}{3}\frac{(r+1)u}{r+u}}{L_{r,u}} = \frac{2/3}{0.07869} > 1,$$
$$\frac{\frac{2r-5}{2r-3}(1 - U_{r,u})}{L_{r,u}} = \frac{0.5801/3}{0.07869} > 1.$$

(2) If $r \geq 4$ and $u = 2$, then $L_{r,u} = \frac{1}{r}$ and $U_{r,u} = \frac{r-1}{r+u}$. Hence

$$\frac{\frac{1}{3}\frac{(r+1)u}{r+u}}{L_{r,u}} = \frac{2(r+1)r}{3(r+2)} \geq \frac{2 \cdot (4+1) \cdot 4}{3 \cdot (4+2)} > 1,$$
$$\frac{\frac{2r-5}{2r-3}(1 - U_{r,u})}{L_{r,u}} = \frac{3(2r-5)r}{(r+2)(2r-3)} \geq \frac{3 \cdot (2 \cdot 4 - 5) \cdot 4}{(4+2)(2 \cdot 4 - 3)} > 1.$$

(3) If $r = 3$ and $u \geq 4$, then $L_{r,u} = \frac{1}{4u}$ and $U_{r,u} = \frac{r-1}{r+u}$. Hence

$$\frac{\frac{1}{3}\frac{(r+1)u}{r+u}}{L_{r,u}} = \frac{16u^2}{3(u+3)} \geq \frac{16 \cdot 4^2}{3 \cdot (4+3)} > 1,$$
$$\frac{\frac{2r-5}{2r-3}(1 - U_{r,u})}{L_{r,u}} = \frac{4u(u+1)}{3(u+3)} \geq \frac{4 \cdot 4 \cdot (4+1)}{3 \cdot (4+3)} > 1.$$

(4) If $r \geq 4$ and $u \geq 3$, then $L_{r,u} = \frac{1}{r+u}$ and $U_{r,u} = \frac{r-1}{r+u}$. Hence

$$\frac{\frac{1}{3}\frac{(r+1)u}{r+u}}{L_{r,u}} = \frac{1}{3}(r+1)u \geq \frac{(4+1)\cdot 3}{3} > 1,$$
$$\frac{\frac{2r-5}{2r-3}(1-U_{r,u})}{L_{r,u}} = \frac{2r-5}{2r-3}(u+1) \geq \frac{2\cdot 4-5}{2\cdot 4-3}\cdot(3+1) > 1. \qquad \square$$

Claim 4.8 *If $r \geq 3$ and $u \geq 2$, but $r \neq 3$ or $u \neq 2$, then*

$$u(u-1)U_{r,u}^r > rL_{r,u}^{r-2}\left[\frac{(r+1)u}{r+u} - L_{r,u}\right](1-U_{r,u}-L_{r,u}).$$

Proof. Let

$$f(r,u) \stackrel{\text{def}}{=} \frac{r}{u(u-1)}L_{r,u}^r\left[\frac{(r+1)u}{r+u} - L_{r,u}\right](1-H_{r,u}-L_{r,u})U_{r,u}^{-r}.$$

We want to show that $f(r,u) < 1$.

(1) If $r = 3$ and $u = 3$, $f(r,u) < 3\cdot 10^{-3} < 1$.

(2) If $r \geq 4$ and $u = 2$,

$$\begin{aligned} f(r,u) &\leq \left[\frac{r+2}{r(r-1)}\right]^3 \cdot \frac{2r^2+r-2}{r+2} \\ &= \left(1+\frac{3}{r-1}\right)^2 \cdot \left(2+\frac{1}{r}-\frac{2}{r^2}\right)\cdot\frac{1}{r(r-1)} < 1. \end{aligned}$$

(3) If $r = 3$ and $u \geq 4$,

$$\begin{aligned} f(r,u) &= \frac{3(u+3)(4u^2+3u-3)(16u^2-u-3)}{512(u-1)u^4} \\ &= \frac{3}{512}\left(1+\frac{4}{u-1}\right)\left[4+3\cdot\frac{1}{u}\left(1-\frac{1}{u}\right)\right] \\ &\quad \cdot\left(16-\frac{1}{u}-\frac{3}{u^2}\right) < 1. \end{aligned}$$

(4) If $r \geq 4$ and $u \geq 3$,

$$\begin{aligned} f(r,u) &= \frac{r}{(r-1)^r}\left(r+1+\frac{r}{u-1}\right) \\ &\leq \frac{r}{(r-1)^4}\left(r+1+\frac{r}{2}\right) \\ &= \frac{1}{2(r-1)^2}\left(1+\frac{1}{r-1}\right)\left(3+\frac{5}{r-1}\right) < 1. \end{aligned}$$ □

A.2 Claims for Quasi-uniform Set-patterns

We will use the following equalities and inequalities in the proof of both Claims 7.1 and 7.2. For any integers $a \geq b \geq 0$ and real numbers $x > y > 0$,

$$\frac{x^a y^b - x^b y^a}{x-y} = \sum_{\substack{i+j=a+b-1 \\ i,j\geq b}} x^i y^j = x^{a-1}y^b + x^{a-2}y^{b+1} + \cdots + x^b y^{a-1} \tag{A.1}$$

$$\leq \frac{a-b}{2}(x^{a-1}y^b + x^b y^{a-1}). \tag{A.2}$$

Particularly, for $b = 0$,

$$\frac{x^a - y^a}{x-y} = \sum_{\substack{i+j=a-1 \\ i\geq 0, j\geq 0}} x^i y^j = x^{a-1} + x^{a-2}y + \cdots + y^{a-1}. \tag{A.3}$$

Claim 7.1 *For any $r_i > r_j > 0$ and $\mu_t \geq 1$,*

$$(r_i r_j + r_i + r_j)^{\mu_t} - (r_i^{\mu_t} + r_j^{\mu_t}) \geq \mu_t (r_i r_j)\left(\frac{r_i^{\mu_t} - r_j^{\mu_t}}{r_i - r_j} + \frac{r_i^{\mu_t - 1} - r_j^{\mu_t - 1}}{r_i - r_j}\right)$$

Proof. Note that

$$\begin{aligned} (r_i r_j + r_i + r_j)^{\mu_t} &\geq \mu_t r_i r_j (r_i + r_j)^{\mu_t - 1} + (r_i + r_j)^{\mu_t} \\ &\geq \mu_t r_i r_j \sum_{\substack{a+b=\mu_t-1 \\ a,b\geq 0}} r_i^a r_j^b + \sum_{\substack{a+b=\mu_t \\ a,b\geq 0}} \mu_t r_i^a r_j^b \\ &= \mu_t r_i r_j \frac{r_i^{\mu_t} - r_j^{\mu_t}}{r_i - r_j} + \mu_t r_i r_j \frac{r_i^{\mu_t - 1} - r_j^{\mu_t - 1}}{r_i - r_j} + (r_i^{\mu_t} + r_j^{\mu_t}), \end{aligned}$$

where the last equality comes from Equality (A.3). The conclusion follows by subtracting both sides by $r_i^{\mu_t} + r_j^{\mu_t}$. □

Claim 7.2 *For any $r_i > r_j > 0$ and $T > \mu_1 \geq \mu_t \geq \mu_{t'} \geq 0$,*

$$\begin{aligned}(\mu_t - 1)\frac{r_i^{\mu_t-1}r_j^{\mu_{t'}} - r_i^{\mu_{t'}}r_j^{\mu_t-1}}{r_i - r_j} \\ + (\mu_{t'} - 1)\frac{r_i^{\mu_{t'}-1}r_j^{\mu_t} - r_i^{\mu_t}r_j^{\mu_{t'}-1}}{r_i - r_j} + (\mu_t - \mu_{t'})\frac{r_i^{\mu_t}r_j^{\mu_{t'}} - r_i^{\mu_{t'}}r_j^{\mu_t}}{r_i - r_j} \\ \leq \left[\frac{T}{T-\mu_1}(\mu_t - \mu_{t'})^2 - (\mu_t + \mu_{t'} - 2)\right](r_i^{\mu_t}r_j^{\mu_{t'}} + r_i^{\mu_{t'}}r_j^{\mu_t})(r_i r_j)^{-1}.\end{aligned}$$

Proof. It is easy to verify that the equality holds if $\mu_t = \mu_{t'}$. Without loss of generality assume $\mu_t > \mu_{t'}$. Then

$$\begin{aligned}\frac{r_i^{\mu_t-1}r_j^{\mu_{t'}} - r_i^{\mu_{t'}}r_j^{\mu_t-1}}{r_i - r_j} &\leq \frac{1}{2}(\mu_t - \mu_{t'} - 1)(r_i^{\mu_t-2}r_j^{\mu_{t'}} + r_i^{\mu_{t'}}r_j^{\mu_t-2}) \\ &\leq \frac{1}{2}(\mu_t - \mu_{t'} - 1)(r_i^{\mu_t-1}r_j^{\mu_{t'}-1} + r_i^{\mu_{t'}-1}r_j^{\mu_t-1}),\end{aligned}$$

and

$$\frac{r_i^{\mu_{t'}-1}r_j^{\mu_t} - r_i^{\mu_t}r_j^{\mu_{t'}-1}}{r_i - r_j} \leq \frac{1}{2}(\mu_t - \mu_{t'} + 1)(r_i^{\mu_t-1}r_j^{\mu_{t'}-1} + r_i^{\mu_{t'}-1}r_j^{\mu_t-1}).$$

Then

$$\begin{aligned}(\mu_t - 1)\frac{r_i^{\mu_t-1}r_j^{\mu_{t'}} - r_i^{\mu_{t'}}r_j^{\mu_t-1}}{r_i - r_j} + (\mu_{t'} - 1)\frac{r_i^{\mu_{t'}-1}r_j^{\mu_t} - r_i^{\mu_t}r_j^{\mu_{t'}-1}}{r_i - r_j} \\ \leq \frac{1}{2}\left[(\mu_t - \mu_{t'})^2 - (\mu_t + \mu_{t'} - 2)\right](r_i^{\mu_t-1}r_j^{\mu_{t'}-1} + r_i^{\mu_{t'}-1}r_j^{\mu_t-1}). \quad \text{(A.4)}\end{aligned}$$

On the other hand, Inequality (A.2) implies that

$$\begin{aligned}(\mu_t - \mu_{t'})\frac{r_i^{\mu_t}r_j^{\mu_{t'}} - r_i^{\mu_{t'}}r_j^{\mu_t}}{r_i - r_j} &\leq \frac{1}{2}(\mu_t - \mu_{t'})^2(r_i^{\mu_t-1}r_j^{\mu_{t'}} + r_i^{\mu_{t'}}r_j^{\mu_t-1}) \\ &\leq \frac{\mu_1}{2(T-\mu_1)}(\mu_t - \mu_{t'})^2(r_i^{\mu_t-1}r_j^{\mu_{t'}-1} + r_i^{\mu_{t'}-1}r_j^{\mu_t-1}), \quad \text{(A.5)}\end{aligned}$$

where we used the majorization property that $p_1 \leq \frac{\mu_1}{T}$ and hence

$$r_j < r_i \leq \frac{\mu_1/T}{1 - \mu_1/T} = \frac{\mu_1}{T - \mu_1}.$$

The conclusion follows by combining Inequalities (A.4) and (A.5). □

A.3 Proof for Almost-uniform Set-patterns

Theorem 7.3 *Given sample time T, the SPML distribution of an almost-uniform set-pattern $\bar{\psi}$ is*

$$\hat{P}_{\bar{\psi},T} = \left(\frac{n}{\hat{k}T}, \ldots, \frac{n}{\hat{k}T}\right)$$

for some $\hat{k} \leq \infty$, where $n = \sum_{t=1}^{m} \mu_t$.

Proof. For simplicity let $P = \hat{P}_{\bar{\psi},T} = (p_1, p_2, \ldots, p_k)$. If there are more than one optimal P, choose one with smallest k.

We have shown in 7.2 that for quasi-uniform set-patterns the SPML is uniform. If $\bar{\psi}$ is not quasi-uniform, *i.e.*, $D_{t,t'} > 0$ for some $\{t, t'\} \in \binom{[m]}{2}$, then the condition in the Theorem implies that $\mu_m > 1$. Thus we may assume that

$$T > \mu_1 > \mu_m > 1.$$

Suppose P is not uniform, *i.e.*, for $p_i > p_j > 0$ for some $i, j \in [k]$. Similar to Inequalities (7.14) and (7.16) in the proof of Theorem 7.2, letting $i = 1$ and $j = k$, we have, for all $\{t, t'\} \in \binom{[m]}{2}$,

$$(\mu_t - 1)\left(r_1^{\mu_t - 1} r_k^{\mu_{t'}} - r_1^{\mu_{t'}} r_k^{\mu_t - 1}\right) + (\mu_{t'} - 1)\left(r_1^{\mu_{t'} - 1} r_k^{\mu_t} - r_1^{\mu_t} r_k^{\mu_{t'} - 1}\right) + (\mu_t - \mu_{t'})\left(r_1^{\mu_t} r_k^{\mu_{t'}} - r_1^{\mu_{t'}} r_k^{\mu_t}\right) \geq 0, \quad \text{(A.6)}$$

and

$$\sum_{\{t,t'\} \in \binom{[m]}{2}} D_{t,t'}(r_1^{\mu_t} r_k^{\mu_{t'}} + r_1^{\mu_{t'}} r_k^{\mu_t}) R_{1,k}(\bar{\psi}_{t,t'}) \geq 0. \quad \text{(A.7)}$$

Let $x = \frac{r_1}{r_k}$. Then, for any $\mu_t \geq \mu_{t'}$, Inequality (A.6) can be written as

$$(\mu_t - 1)(x^{\mu_t - 1} - x^{\mu_{t'}}) + (\mu_{t'} - 1)(x^{\mu_{t'} - 1} - x^{\mu_t}) + r_1(\mu_t - \mu_{t'})(x^{\mu_t - 1} - x^{\mu_{t'}}) \geq 0.$$

It follows that

$$\begin{aligned}(\mu_{t'} - 1)x^{\mu_t} &\leq \left[(\mu_t - 1) + r_1(\mu_t - \mu_{t'})\right] x^{\mu_t - 1} + (\mu_{t'} - 1)x^{\mu_{t'} - 1} \\ &\quad - \left[(\mu_t - 1) + r_1(\mu_t - \mu_{t'})\right] x^{\mu_{t'}} \\ &\leq \left[(\mu_t - 1) + r_1(\mu_t - \mu_{t'})\right] x^{\mu_t - 1}.\end{aligned}$$

Thus

$$x = \frac{r_1}{r_k} \leq \frac{(\mu_t - 1) + r_1(\mu_t - \mu_{t'})}{\mu_{t'} - 1} \leq \frac{(\mu_1 - 1) + \frac{\mu_1}{T-\mu_1}(\mu_1 - \mu_m)}{\mu_m - 1}. \tag{A.8}$$

On the other hand, note that in Inequality (A.7), for any $\mu_t \geq \mu_{t'}$, $R_{1,k}(\bar{\psi}_{t,t'})$ is a polynomial in $r_2, r_3, \ldots, r_{k-1}$. Then

$$\begin{aligned}
\frac{(r_1^{\mu_t} r_k^{\mu_{t'}} + r_1^{\mu_{t'}} r_k^{\mu_t}) R_{1,k}(\bar{\psi}_{t,t'})}{(r_1^{\mu_1} r_k^{\mu_m} + r_1^{\mu_m} r_k^{\mu_1}) R_{1,k}(\bar{\psi}_{1,m})} &\geq \min_{i_1,i_2} \frac{(r_1^{\mu_t} r_k^{\mu_{t'}} + r_1^{\mu_{t'}} r_k^{\mu_t}) r_{i_1}^{\mu_1} r_{i_2}^{\mu_m}}{(r_1^{\mu_1} r_k^{\mu_m} + r_1^{\mu_m} r_k^{\mu_1}) r_{i_1}^{\mu_t} r_{i_2}^{\mu_{t'}}} \\
&\geq \frac{2 r_1^{\mu_{t'}} r_k^{\mu_t} \cdot r_k^{\mu_1} r_1^{\mu_m}}{2 r_1^{\mu_1} r_k^{\mu_m} \cdot r_k^{\mu_t} r_1^{\mu_{t'}}} \\
&= \left(\frac{r_k}{r_1}\right)^{\mu_1 - \mu_m}.
\end{aligned}$$

Similarly

$$\begin{aligned}
\frac{(r_1^{\mu_t} r_k^{\mu_{t'}} + r_1^{\mu_{t'}} r_k^{\mu_t}) R_{1,k}(\bar{\psi}_{t,t'})}{(r_1^{\mu_1} r_k^{\mu_m} + r_1^{\mu_m} r_k^{\mu_1}) R_{1,k}(\bar{\psi}_{1,m})} &\leq \max_{i_1,i_2} \frac{(r_1^{\mu_t} r_k^{\mu_{t'}} + r_1^{\mu_{t'}} r_k^{\mu_t}) r_{i_1}^{\mu_1} r_{i_2}^{\mu_m}}{(r_1^{\mu_1} r_k^{\mu_m} + r_1^{\mu_m} r_k^{\mu_1}) r_{i_1}^{\mu_t} r_{i_2}^{\mu_{t'}}} \\
&\leq \frac{1 + \left(\frac{r_k}{r_1}\right)^{\mu_t - \mu_{t'}}}{1 + \left(\frac{r_k}{r_1}\right)^{\mu_1 - \mu_m}} \cdot \frac{r_1^{\mu_t} r_k^{\mu_{t'}} \cdot r_1^{\mu_1} r_k^{\mu_m}}{r_1^{\mu_1} r_k^{\mu_m} \cdot r_1^{\mu_t} r_k^{\mu_{t'}}} \\
&\leq 1,
\end{aligned}$$

Then Inequality (A.6) implies that

$$\sum_{D_{t,t'}>0} D_{t,t'} \geq \sum_{D_{t,t'}<0} |D_{t,t'}| \left(\frac{r_k}{r_1}\right)^{\mu_1 - \mu_m}. \tag{A.9}$$

Combining Inequalities (A.8) and (A.9) we get

$$\sum_{D_{t,t'}>0} D_{t,t'} \left[\frac{(\mu_1 - 1) + \frac{\mu_1}{T-\mu_1}(\mu_1 - \mu_m)}{\mu_m - 1}\right]^{\mu_1 - \mu_m} > \sum_{D_{t,t'}<0} |D_{t,t'}|,$$

which contradicts the condition given in the theorem. □

Index

Bibliography

[ADJ+11] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *IEEE International Symposium on Information Theory*, 2011.

[ADM+10] Jayadev Acharya, Hirakendu Das, Hosein Mohimani, Alon Orlitsky, and Shengjun Pan. Exact calculation of pattern probabilities. In *ISIT'10: Proceedings of the 2010 IEEE international conference on Symposium on Information Theory*, 2010.

[AOP09] Jayadev Acharya, Alon Orlitsky, and Shengjun Pan. The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns. In *IEEE International Symposium on Information Theory*, 2009.

[BF93] J. Bunge and M. Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88:364–373, 1993.

[Com74] L Comtet. *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Springer, 1974.

[DO06] Anand K. Dhulipala and Alon Orlitsky. Universal compression of Markov and related sources over arbitrary alphabets. *IEEE Transactions on Information Theory*, 52(9):4182 – 4190, 2006.

[ET76] B. Efron and R. Thisted. Estimating the number of unseen species: How many words did Shakespeare know. *Biometrika*, 63:435–447, 1976.

[FCW43] R. Fisher, A. Corbet, and C. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12:42–48, 1943.

[GT56] I.J. Good and G.H. Toulmin. The number of new species and the increase in population coverage when the sample is increased. *Biometrika*, 43(1):45–63, 1956.

[KET06] Jon Kleinberg and Éva Tardos. *Algorithm Design*. Addison-Wesley, 2006.

[MR68] C. L. Mallows and John Riordan. The inversion enumerator for labeled trees. *Bull. Amer. Math. Soc.*, 74(1):92–94, 1968.

[Mui02] R. F. Muirhead. Some methods applicable to identities and inequalities of symmetric algebraic functions of n letters. In *Proceedings of the Edinburgh Mathematical Society*, volume 21, pages 144–162, 1902.

[OP09] Alon Orlitsky and Shengjun Pan. The maximum likelihood probability of skewed patterns. In *IEEE International Symposium on Information Theory*, 2009.

[OPS+12] A. Orlitsky, Shengjun Pan, Sajama, N.P. Santhanam, K. Viswanathan, and J. Zhang. Pattern maximum likelihood: computation and experiments. In preparation, 2012.

[OSS+04] A. Orlitsky, Sajama, N.P. Santhanam, K. Viswanathan, and J. Zhang. Algorithms for modeling distributions over large alphabets. In *ISIT'04: Proceedings of the 2004 IEEE international conference on Symposium on Information Theory*, 2004.

[OSVZ04] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004.

[OSVZ12] A. Orlitsky, N.P. Santhanam, K. Viswanathan, and J. Zhang. Pattern maximum likelihood: existence and properties. In preparation, 2012.

[OSZ04] A. Orlitsky, N.P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50(7):1469 – 1481, 2004.

[Rob55] H. Robbins. A remark on stirlings formula. *Amer. Math. Monthly*, pages 26–29, 1955.

[TE87] R. Thisted and B. Efron. Did shakespeare write a newly-discovered poem? *Biometrika*, 74:445–455, 1987.

[Val08] Paul Valiant. Testing symmetric properties of distributions. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 383–392, 2008.

[vLW92] J. H. van Lint and R. M. Wilson. *A course in combinatorics*. Cambridge University Press, Cambridge, 1992.

[VV11a] Gregory Valiant and Paul Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new clts. In *roceedings of the 43rd ACM Symposium on Theory of Computing*, pages 685–694, 2011.

[VV11b] Gregory Valiant and Paul Valiant. The power of linear estimators. In *IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 403–412, 2011.

[Zha05] Junan Zhang. *Universal Compression and Probability Estimation with Unknown Alphabets*. PhD thesis, University of California, San Diego, 2005.