Department of Computer Science
Series of Publications A
Report A-2009-11

Computationally Efficient Methods for MDL-Optimal Density Estimation and Data Clustering

Petri Kontkanen

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium CK112, Exactum on November 30th, 2009, at 12 o'clock noon.

University of Helsinki
Finland

Contact information

Postal address:
Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: postmaster@cs.Helsinki.FI (Internet)

URL: http://www.cs.Helsinki.FI/

Telephone: +358 9 1911

Telefax: +358 9 191 51120

Copyright © 2009 Petri Kontkanen
ISSN 1238-8645
ISBN 978-952-10-5900-1 (paperback)
ISBN 978-952-10-5901-8 (PDF)
Computing Reviews (1998) Classification: G.2.1, G.3, H.1.1
Helsinki 2009
Helsinki University Print

Computationally Efficient Methods for MDL-Optimal Density Estimation and Data Clustering

Petri Kontkanen

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
Petri.Kontkanen@cs.Helsinki.FI
http://www.cs.helsinki.fi/u/pkontkan/

PhD Thesis, Series of Publications A, Report A-2009-11
Helsinki, November 2009, 75+64 pages
ISSN 1238-8645
ISBN 978-952-10-5900-1 (paperback)
ISBN 978-952-10-5901-8 (PDF)

Abstract

The Minimum Description Length (MDL) principle is a general, well-founded theoretical formalization of statistical modeling. The most important notion of MDL is the stochastic complexity, which can be interpreted as the shortest description length of a given sample of data relative to a model class. The exact definition of the stochastic complexity has gone through several evolutionary steps. The latest instantation is based on the so-called Normalized Maximum Likelihood (NML) distribution which has been shown to possess several important theoretical properties. However, the applications of this modern version of the MDL have been quite rare because of computational complexity problems, i.e., for discrete data, the definition of NML involves an exponential sum, and in the case of continuous data, a multi-dimensional integral usually infeasible to evaluate or even approximate accurately. In this doctoral dissertation, we present mathematical techniques for computing NML efficiently for some model families involving discrete data. We also show how these techniques can be used to apply MDL in two practical applications: histogram density estimation and clustering of multi-dimensional data.

Computing Reviews (1998) Categories and Subject Descriptors:
G.2.1 [Combinatorics]
G.3 [Probability and Statistics]

H.1.1 [Models and Principles]: Systems and Information theory

General Terms:
statistics, machine learning, algorithms

Additional Key Words and Phrases:
information theory, minimum description length, density estimation, clustering

Preface

This doctoral dissertation consists of an introductory part and six original research papers on the Minimum Description Length (MDL) principle. The focus of the papers is on the practical aspects of the MDL, not the theoretical properties of it. More precisely, the research papers present mathematical techniques that allow the efficient use of MDL in practical model class selection tasks. The papers also discuss how these techniques can be applied in real-world applications.

To give the reader preliminaries and motivation for easier understanding of the six research papers, the thesis starts with an introductory text. This part is intuitive in nature, all the technical details can be found in the respective research papers. The introductory part starts with a short review of the MDL principle and the NML distribution, which formally defines the MDL model class selection criterion (the stochastic complexity). Next, an overview of the mathematical techniques and algorithms for efficient computation of the NML is presented. These algorithms are then used in two practical applications: histogram density estimation and clustering of multi-dimensional data.

The final part of the introduction consists of two appendices. The first one provides the reader background to the mathematical tools used in various parts of the thesis. The topics of this appendix are complex analysis, formal power series, generating functions and asymptotic analysis of generating functions. Together these techniques provide a powerful toolbox for efficient NML computation for several interesting model families. The topic of the second appendix is the derivation of a novel, very accurate multinomial NML approximation. The derivation is based on the mathematical techniques described in the first appendix.

Acknowledgements

I would like to thank my advisor, Professor Petri Myllymäki, for several invaluable discussions and comments regarding this dissertation. I am also very grateful to Henry Tirri and Petri for providing me an inspiring and fun working environment in the CoSCo research group for so many years.

The Department of Computer Science of the University of Helsinki provided me a chance for a research visit to the Hong Kong University of Science and Technology, where the introductory part of this thesis was written. I want to thank my host Professor Nevin Zhang and his research group for many useful discussions about various issues related to my research work. They also helped me a lot in adjusting to life and culture in Hong Kong.

In addition to the Department of Computer Science, the support from the Helsinki Institute for Information Technology (HIIT), the Academy of Finland, the PASCAL EU Network of Excellence, the Finnish Funding Agency for Technology and Innovation (Tekes) and the Helsinki Graduate School in Computer Science and Engineering (Hecse) have made it possible to conduct the research work of my dissertation.

I am also very grateful to my long time colleagues of the CoSCo research group, including Henry Tirri, Petri Myllymäki, Jussi Lahtinen, Tomi Silander, Tommi Mononen, Hannes Wettig, Antti Tuominen, Jukka Perkiö, Kimmo Valtonen and Teemu Roos. In addition, I want to thank the project secretary of CoSCo, Taina Nikko, and the HIIT IT group, lead by Pekka Tonteri.

The pre-examiners of the manuscript of this dissertation were Professors Nevin Zhang and Samuel Kaski. I want to thank them for their effort and positive comments.

Finally, I want to thank my family and friends for all of their support and encouragement throughout the process of writing my dissertation.

Original publications and contributions

This doctoral dissertation is based on the following 6 research papers, which are referred in text as Papers I–VI. The papers are re-printed at the end of the thesis.

Paper I: P. Kontkanen, W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri. Efficient computation of stochastic complexity. In C. Bishop and B. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2003.

Paper II: P. Kontkanen and P. Myllymäki. A fast normalized maximum likelihood algorithm for multinomial data. In L. P. Kaelbling and A. Saffiotti, editors, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, 2005.

Paper III: P. Kontkanen and P. Myllymäki. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103(6):227–233, 2007.

Paper IV: P. Kontkanen and P. Myllymäki. MDL histogram density estimation. In M. Meila and S. Shen, editors, *Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics*, March 2007.

Paper V: P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri. An MDL framework for data clustering. In P. Grünwald, I. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*. The MIT Press, 2005.

Paper VI: P. Kontkanen and P. Myllymäki. An empirical comparison of NML clustering algorithms. In M. Dehmer, M. Drmota, and F. Emmert-Streib, editors, *Proceedings of the International Conference on Information Theory and Statistical Learning (ITSL-08)*. CSREA Press, 2008.

In Papers I-III we develop algorithms for efficient computation of the NML in the case of the multinomial and Naive Bayes model family. The topic of Papers IV–VI is to show how NML can be applied to practical problems. The main contributions and short descriptions of the six papers are listed here:

Paper I: We introduce the first polynomial-time algorithm for computing the stochastic complexity (NML) for the multinomial and Naive Bayes model families. The running time of the algorithm is quadratic with respect to the sample size. We also present three stochastic complexity approximation algorithms and study their accuracy empirically.

Paper II: We improve the time complexity of the algorithm presented in Paper I to $\mathcal{O}(n \log n)$, where n is the sample size. The new algorithm is based on the convolution theorem and the Fast Fourier Transform (FFT) algorithm.

Paper III: We derive a recursion formula that can be used straightforwardly to compute the multinomial stochastic complexity in linear time. The mathematical technique applied here is generating functions.

Paper IV: We regard histogram density estimation as a model class selection problem and apply the minimum description length (MDL) principle to it. Using the results from Paper III, we show how to efficiently compute the stochastic complexity for the histogram densities. Furthermore, we derive a dynamic programming algorithm that can be used to find the globally optimal histogram in polynomial time.

Paper V: Clustering is one of the central concepts in the field of unsupervised data analysis. We regard clustering as a problem of partitioning the data into mutually exclusive clusters so that similar data vectors are grouped together. The number of clusters is unknown, and determining the optimal number is part of the clustering problem. For solving this problem, we suggest an information-theoretic framework based on the MDL principle. For computing the NML for the clustering model class, we use the algorithms of Papers I and II.

Paper VI: We compare empirically various algorithms for finding candidate solutions to the clustering problem discussed in Paper V. We present five algorithms for the task and use several real-world data sets to test the

algorithms. The results show that the traditional EM and K-means algorithms perform poorly. Furthermore, our novel hybrid clustering algorithm turns out to produce the best results.

In all six papers, the contribution of the current author is significant. In Paper I, the quadratic-time algorithm for the multinomial model family is due to Wray Buntine. The idea of applying MDL to the clustering problem in Paper V is by Petri Myllymäki.

Contents

Preface **v**

Acknowledgements **vii**

Original publications and contributions **ix**

1 Introduction **1**

2 Stochastic Complexity **5**
2.1 Model Classes and Families 5
2.2 The Normalized Maximum Likelihood (NML) Distribution . 6
2.3 NML for the Multinomial Model Family 7
2.4 NML for the Naive Bayes Model Family 8

3 Efficient Computation of NML **11**
3.1 The Multinomial Model Family 11
3.1.1 Exact Computation Algorithms 11
3.1.2 NML Approximations 13
3.1.3 Comparison of the Approximations 15
3.2 The Naive Bayes Model Family 19

4 MDL Applications **21**
4.1 Histogram Density Estimation 22
4.1.1 Definitions 22
4.1.2 NML Histogram 23
4.2 Clustering 26
4.2.1 NML Clustering 26
4.2.2 Comparison of Clustering Algorithms 27

5 Conclusion **31**

Appendices **35**

A Mathematical Background **35**
A.1 Review of Complex Analysis 35
A.1.1 The Complex Numbers and the Complex Plane . . . 36
A.1.2 Roots of Complex Numbers 37
A.1.3 Analytic Functions 38
A.1.4 Complex Integration 39
A.1.5 Laurent Expansion 39
A.1.6 The Residue Theorem 40
A.1.7 Puiseux Expansion 41
A.2 Formal Power Series 42
A.2.1 Definition 42
A.2.2 Linear Combination 42
A.2.3 Multiplication 43
A.2.4 Reciprocal Series 43
A.2.5 Inverse Series 44
A.3 Generating Functions 45
A.3.1 Definition 45
A.3.2 Fibonacci Numbers 46
A.3.3 Integer Partitions 47
A.4 Asymptotic Analysis of Generating Functions 49
A.4.1 Rational Functions 49
A.4.2 Asymptotics of Integer Partitions 52
A.4.3 Algebraic-Logarithmic Functions: The Singularity Analysis 52

B The Szpankowski Approximation **57**
B.1 The Regret Generating Function 57
B.2 The Derivation 59

References **69**

Chapter 1

Introduction

Many problems in science can be cast as *model class selection* tasks, i.e., as tasks of selecting among a set of competing mathematical explanations the one that describes a given sample of data best. The *Minimum description length* (MDL) principle developed in the series of papers [53, 54, 56] is a well-founded, general framework for performing model class selection and other types of statistical inference. The fundamental idea behind the MDL principle is that any regularity in data can be used to *compress* the data, i.e., to find a description or *code* of it, such that this description uses less symbols than it takes to describe the data literally. The more regularities there are, the more the data can be compressed. According to the MDL principle, learning can be equated with finding regularities in data. Consequently, we can say that the more we are able to compress the data, the more we have learned about them.

The MDL principle has several desirable properties. Firstly, it automatically protects against overfitting in the model class selection process. Secondly, this statistical framework does not – unlike most other frameworks – assume that there exists some underlying "true" model. The model class is only used as a technical device for constructing an efficient code for describing the data. MDL is also closely related to the Bayesian inference but there are some fundamental differences, the most important being that MDL does not need any prior distribution; it only uses the data at hand. For more discussion on the theoretical motivations behind the MDL principle see, e.g., [56, 5, 72, 57, 21, 58].

MDL model class selection is based on a quantity called *stochastic complexity*, which is the shortest description length of a given data relative to a model class. The stochastic complexity is defined via the normalized maximum likelihood (NML) distribution [63, 56]. For multinomial (discrete) data, this definition involves a normalizing sum over all the possible data

samples of a fixed size. The logarithm of this sum is called the *parametric complexity* or *regret*, which can be interpreted as the amount of complexity of the model class. If the data is continuous, the sum is replaced by the corresponding integral.

The NML distribution has several theoretical optimality properties, which make it a very attractive candidate for performing model class selection and related tasks. It was originally [56, 5] formulated as the unique solution to a minimax problem presented in [63], which implied that NML is the minimax optimal universal model. Later [57], it was shown that NML is also the solution to a related problem involving expected regret. See Section 2.2 and [5, 57, 21, 58] for more discussion on the theoretical properties of the NML.

Many scientific problems involve large data sets. In order to apply NML for these tasks one needs to develop suitable NML computation methods since the normalizing sum or integral in the definition of the NML is typically difficult to compute directly. The introductory part of this thesis starts by presenting algorithms for efficient computation of the NML for both one- and multi-dimensional discrete data. The model families used here are the multinomial and the Naive Bayes, and the discussion is based on the Papers I–III. In the multinomial case, the most efficient algorithm based on the technique of generating functions is linear with respect to the sample size, while the Naive Bayes algorithm is quadratic.

The task of finding efficient NML computation algorithms is a relatively new topic, and there are only few related studies. In [50], NML for the multinomial model family was written in another form, which resulted in another linear-time algorithm. The same paper also studied the connection between the multinomial NML and the so-called *birthday problem* [15], which is a classical problem of probability theory. A study of how the multinomial NML can be computed in sub-linear time with a finite precision is presented in [47]. The algorithm has time complexity $\mathcal{O}(\sqrt{dn})$, where d is the precision in digits and n is the sample size. In [49], new theoretically interesting recurrence formulas for NML computation are derived. A new quadratic-time algorithm for computing the parametric complexity in the case of Naive Bayes is presented in [46]. This algorithm is based on the so-called *Miller formula* [25] for computing the powers of formal power series.

There has also been studies on computing NML for more complex model families. In [70, 42, 48], algorithms for so-called *Bayesian forests* are presented. However, these algorithms are exponential with respect to the number of values of the domain variables. One solution to this problem

is suggested in [61], where the NML criterion is modified to a computationally less demanding form called the *factorized NML*. Initial empirical results show that this new criterion can be useful in model class selection problems.

The second part of the thesis describes how NML can be applied to practical problems using the techniques of the first part. Due to the computational efficiency problems, there are relatively few applications of NML. However, the existing applications demonstrate that NML works very well in practice and provides in many cases superior results when compared to alternative approaches. The first application discussed in the thesis is the NML-optimal histogram density estimation suggested in Paper IV. This framework provides both the optimal number of bins and the location of the bin borders of the histogram in polynomial time. The second application is the NML clustering of multi-dimensional discrete data introduced in Paper V. The optimization aspect of the clustering problem was studied in Paper VI, where five algorithms for efficiently searching the exponentially-sized clustering space were compared. See Chapter 4 for related work and more discussion on NML applications in general.

This thesis is structured as follows. In Chapter 2 we discuss the basic properties of the MDL principle and the NML distribution. We also instantiate NML for the two model families. In Chapter 3 we present both exact and approximative computation algorithms for NML. The chapter also includes an empirical comparison of three NML approximations for the multinomial model family. The topic of Chapter 4 is to show how NML can be applied in two practical tasks: density estimation and data clustering. Chapter 5 gives some concluding remarks and ideas for future work. The thesis then continues with two appendices: Appendix A provides mathematical background to the techniques used in the thesis and Appendix B gives a full derivation of the accurate multinomial NML approximation called the Szpankowski approximation. Finally, the six original research papers are re-printed at the end of the thesis.

Chapter 2

Stochastic Complexity

The MDL model class selection is based on minimization of the stochastic complexity. In the following, we first define the model class selection problem. Then we proceed by giving the definition of the stochastic complexity based on the normalized maximum likehood distribution and discuss its theoretical properties. Finally, we instantiate the NML for the multinomial and Naive Bayes model families.

2.1 Model Classes and Families

Let $\mathbf{x}^n = (x_1, \ldots, x_n)$ be a data sample of n outcomes, where each outcome x_j is an element of some space of observations $\mathcal{X}$. The n-fold Cartesian product $\mathcal{X} \times \cdots \times \mathcal{X}$ is denoted by $\mathcal{X}^n$, so that $\mathbf{x}^n \in \mathcal{X}^n$. Consider a set $\Theta \subseteq \mathbb{R}^d$, where d is a positive integer. A class of parametric distributions indexed by the elements of Θ is called a *model class*. That is, a model class $\mathcal{M}$ is defined as

$$\mathcal{M} = \{P(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}, \tag{2.1}$$

and the set Θ is called the *parameter space*.

Consider now a set $\Phi \subseteq \mathbb{R}^e$, where e is a positive integer. Define a set $\mathcal{F}$ by

$$\mathcal{F} = \{\mathcal{M}(\boldsymbol{\phi}) : \boldsymbol{\phi} \in \Phi\}. \tag{2.2}$$

The set $\mathcal{F}$ is called a *model family*, and each of the elements $\mathcal{M}(\boldsymbol{\phi})$ is a model class. The associated parameter space is denoted by $\Theta_{\boldsymbol{\phi}}$. The model class selection problem can now be defined as a process of finding the parameter vector $\boldsymbol{\phi}$, which is optimal according to some pre-determined criteria. In Sections 2.3 – 2.4 we discuss two specific model families, which will make these definitions more concrete.

2.2 The Normalized Maximum Likelihood (NML) Distribution

One of the most theoretically and intuitively appealing model class selection criteria is the *stochastic complexity*. Denote first the maximum likelihood estimate of data $\mathbf{x}^n$ for a given model class $\mathcal{M}(\boldsymbol{\phi})$ by $\hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(\boldsymbol{\phi}))$, i.e., $\hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(\boldsymbol{\phi})) = \arg\max_{\boldsymbol{\theta} \in \Theta_{\boldsymbol{\phi}}} \{P(\mathbf{x}^n \mid \boldsymbol{\theta})\}$. The *normalized maximum likelihood* (NML) distribution [63] is now defined as

$$P_{\mathrm{NML}}(\mathbf{x}^n \mid \mathcal{M}(\boldsymbol{\phi})) = \frac{P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(\boldsymbol{\phi})))}{\mathcal{C}(\mathcal{M}(\boldsymbol{\phi}), n)}, \tag{2.3}$$

where the normalizing term $\mathcal{C}(\mathcal{M}(\boldsymbol{\phi}), n)$ in the case of discrete data is given by

$$\mathcal{C}(\mathcal{M}(\boldsymbol{\phi}), n) = \sum_{\mathbf{y}^n \in \mathcal{X}^n} P(\mathbf{y}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{y}^n, \mathcal{M}(\boldsymbol{\phi}))), \tag{2.4}$$

and the sum goes over the space of data samples of size n. If the data is continuous, the sum is replaced by the corresponding integral.

The stochastic complexity of the data $\mathbf{x}^n$, given a model class $\mathcal{M}(\boldsymbol{\phi})$, is defined via the NML distribution as

$$\mathrm{SC}(\mathbf{x}^n \mid \mathcal{M}(\boldsymbol{\phi})) = -\log P_{\mathrm{NML}}(\mathbf{x}^n \mid \mathcal{M}(\boldsymbol{\phi})) \tag{2.5}$$

$$= -\log P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(\boldsymbol{\phi}))) + \log \mathcal{C}(\mathcal{M}(\boldsymbol{\phi}), n), \tag{2.6}$$

and the term $\log \mathcal{C}(\mathcal{M}(\boldsymbol{\phi}), n)$ is called the *(minimax) regret* or *parametric complexity*. The regret can be interpreted as measuring the logarithm of the number of essentially different (distinguishable) distributions in the model class. Intuitively, if two distributions assign high likelihood to the same data samples, they do not contribute much to the overall complexity of the model class, and the distributions should not be counted as different for the purposes of statistical inference. See [4] for more discussion on this topic.

The NML distribution (2.3) has several important theoretical optimality properties. The most important one is that NML provides a unique solution to the minimax problem posed in [63]:

$$\min_{\hat{P}} \max_{\mathbf{x}^n} \log \frac{P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(\boldsymbol{\phi})))}{\hat{P}(\mathbf{x}^n \mid \mathcal{M}(\boldsymbol{\phi}))}, \tag{2.7}$$

where $\hat{P}$ can be any distribution over the data $\mathbf{x}^n$. The minimizing $\hat{P}$ is the NML distribution, and the minimax regret

$$\log P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(\boldsymbol{\phi}))) - \log \hat{P}(\mathbf{x}^n \mid \mathcal{M}(\boldsymbol{\phi})) \tag{2.8}$$

is given by the parametric complexity $\log \mathcal{C}(\mathcal{M}(\boldsymbol{\phi}), n)$. This means that the NML distribution is the *minimax optimal universal model*. The term universal model in this context means that the NML distribution represents (or mimics) the behaviour of all the distributions in the model class $\mathcal{M}(\boldsymbol{\phi})$. Note that the NML distribution itself does not have to belong to the model class, and typically it does not. For more discussion on the theoretical properties of NML, see [5, 57, 21, 58].

2.3 NML for the Multinomial Model Family

In the case of discrete data, the simplest model family is the *multinomial*. The data is assumed to be one-dimensional and have only a finite set of possible values. Although simple, the multinomial model family has practical applications. In Paper IV, multinomial NML was used for histogram density estimation, and the problem was regarded as a model class selection task. The NML-optimal histograms were later [12] used as attribute models for Naive Bayes classifier.

Assume that our problem domain consists of a single discrete random variable X with K values, and that our data $\mathbf{x}^n = (x_1, \ldots, x_n)$ is multinomially distributed. The space of observations $\mathcal{X}$ is now the set $\{1, 2, \ldots, K\}$. The corresponding model family $\mathcal{F}_{\text{MN}}$ is defined by

$$\mathcal{F}_{\text{MN}} = \{\mathcal{M}(\boldsymbol{\phi}) : \boldsymbol{\phi} \in \Phi_{\text{MN}}\}, \tag{2.9}$$

where $\Phi_{\text{MN}} = \{1, 2, 3, \ldots\}$. Since the parameter vector $\boldsymbol{\phi}$ is in this case a single integer K, we denote the multinomial model classes by $\mathcal{M}(K)$ for simplicity and define

$$\mathcal{M}(K) = \{P(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_K\}, \tag{2.10}$$

where Θ_K is the simplex-shaped parameter space

$$\Theta_K = \{(\pi_1, \ldots, \pi_K) : \pi_k \geq 0,\ \pi_1 + \cdots + \pi_K = 1\}, \tag{2.11}$$

with $\pi_k = P(X = k),\ k = 1, \ldots, K$.

Assume the data points x_j are independent and identically distributed (i.i.d.). The NML distribution (2.3) for the model class $\mathcal{M}(K)$ is now given by (see Papers I and V)

$$P_{\text{NML}}(\mathbf{x}^n \mid \mathcal{M}(K)) = \frac{\prod_{k=1}^{K} \left(\frac{h_k}{n}\right)^{h_k}}{\mathcal{C}(\mathcal{M}(K), n)}, \tag{2.12}$$

where h_k is the frequency (number of occurrences) of value k in $\mathbf{x}^n$, and

$$\mathcal{C}(\mathcal{M}(K), n) = \sum_{\mathbf{y}^n} P(\mathbf{y}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{y}^n, \mathcal{M}(K))) \tag{2.13}$$

$$= \sum_{h_1+\cdots+h_K=n} \frac{n!}{h_1! \cdots h_K!} \prod_{k=1}^{K} \left(\frac{h_k}{n}\right)^{h_k}. \tag{2.14}$$

2.4 NML for the Naive Bayes Model Family

The one-dimensional case discussed in the previous section is not adequate for many real-world situations, where data are typically multi-dimensional, involving complex dependencies between the domain variables. In Paper I a quadratic-time algorithm for computing the NML for a specific multivariate model family, usually called the Naive Bayes, was derived. This model family has been very successful in practice in mixture modeling [41], clustering of data (Paper V), case-based reasoning [39], classification [22, 40] and data visualization [33].

Let us assume that our problem domain consists of m primary variables $X_1, \ldots, X_m$ and a special variable X_0, which can be one of the variables in our original problem domain or it can be latent. Assume that the variable X_i has K_i values and that the extra variable X_0 has K_0 values. The data $\mathbf{x}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ consist of observations of the form $\mathbf{x}_j = (x_{j0}, x_{j1}, \ldots, x_{jm}) \in \mathcal{X}$, where

$$\mathcal{X} = \{1, 2, \ldots, K_0\} \times \{1, 2, \ldots, K_1\} \times \cdots \times \{1, 2, \ldots, K_m\}. \tag{2.15}$$

The Naive Bayes model family $\mathcal{F}_{\mathrm{NB}}$ is defined by

$$\mathcal{F}_{\mathrm{NB}} = \{\mathcal{M}(\boldsymbol{\phi}) : \boldsymbol{\phi} \in \Phi_{\mathrm{NB}}\} \tag{2.16}$$

with $\Phi_{\mathrm{NB}} = \{1, 2, 3, \ldots\}^{m+1}$. The corresponding model classes are denoted by $\mathcal{M}(K_0, K_1, \ldots, K_m)$:

$$\mathcal{M}(K_0, K_1, \ldots, K_m) = \{P_{\mathrm{NB}}(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{K_0, K_1, \ldots, K_m}\}. \tag{2.17}$$

The basic Naive Bayes assumption is that given the value of the special variable, the primary variables are independent. We have consequently

$$P_{\mathrm{NB}}(X_0 = x_0, X_1 = x_1, \ldots, X_m = x_m \mid \boldsymbol{\theta}) = P(X_0 = x_0 \mid \boldsymbol{\theta}) \cdot \prod_{i=1}^{m} P(X_i = x_i \mid X_0 = x_0, \boldsymbol{\theta}). \tag{2.18}$$

Furthermore, we assume that the distribution of $P(X_0 \mid \boldsymbol{\theta})$ is multinomial with parameters $(\pi_1, \ldots, \pi_{K_0})$, and each $P(X_i \mid X_0 = k, \boldsymbol{\theta})$ is multinomial with parameters $(\sigma_{ik1}, \ldots, \sigma_{ikK_i})$. The whole parameter space is then

$$\begin{aligned}\Theta_{K_0,K_1,\ldots,K_m} =&\{(\pi_1, \ldots, \pi_{K_0}), (\sigma_{111}, \ldots, \sigma_{11K_1}), \ldots, (\sigma_{mK_01}, \ldots, \sigma_{mK_0K_m}) : \\ &\pi_k \geq 0,\ \sigma_{ikl} \geq 0,\ \pi_1 + \cdots + \pi_{K_0} = 1,\ \sigma_{ik1} + \cdots + \sigma_{ikK_i} = 1, \\ &i = 1, \ldots, m,\ k = 1, \ldots K_0\}, \end{aligned} \tag{2.19}$$

and the parameters have interpretations $\pi_k = P(X_0 = k)$ and $\sigma_{ikl} = P(X_i = l \mid X_0 = k)$.

Assuming i.i.d., the NML distribution for the Naive Bayes can now be written as (see Paper V)

$$P_{\mathrm{NML}}(\mathbf{x}^n \mid \mathcal{M}(K_0, K_1, \ldots, K_m)) = \frac{\prod_{k=1}^{K_0} \left(\frac{h_k}{n}\right)^{h_k} \prod_{i=1}^{m} \prod_{l=1}^{K_i} \left(\frac{f_{ikl}}{h_k}\right)^{f_{ikl}}}{\mathcal{C}(\mathcal{M}(K_0, K_1, \ldots, K_m), n)}, \tag{2.20}$$

where h_k is the number of times X_0 has value k in $\mathbf{x}^n$, f_{ikl} is the number of times X_i has value l when the special variable has value k, and $\mathcal{C}(\mathcal{M}(K_0, K_1, \ldots, K_m), n)$ is given by

$$\begin{aligned}&\mathcal{C}(\mathcal{M}(K_0, K_1, \ldots, K_m), n) \\ &\qquad = \sum_{h_1 + \cdots + h_{K_0} = n} \frac{n!}{h_1! \cdots h_{K_0}!} \prod_{k=1}^{K_0} \left(\frac{h_k}{n}\right)^{h_k} \prod_{i=1}^{m} \mathcal{C}(\mathcal{M}(K_i), h_k). \end{aligned} \tag{2.21}$$

Chapter 3

Efficient Computation of NML

In the previous chapter we saw that in the case of discrete data the definition of the NML distribution involves a sum over all the possible data samples of fixed size. Direct computation of this sum takes exponential time even in the case of a simple multinomial model. In this chapter we present efficient algorithms for computing this sum for two model families, the multinomial and Naive Bayes. For interesting algorithms for computing the NML for a more complex model family called the *Bayesian forests*, see [70, 42, 48].

3.1 The Multinomial Model Family

3.1.1 Exact Computation Algorithms

In the previous chapter we saw that the NML distribution for the multinomial model family (2.12) consists of two parts: the likelihood and the parametric complexity (2.14). It is clear that the likelihood term can be computed in linear time by simply sweeping through the data once and counting the frequencies h_k. However, the normalizing sum $\mathcal{C}(\mathcal{M}(K), n)$ (and thus also the parametric complexity $\log \mathcal{C}(\mathcal{M}(K), n)$) involves a sum over an exponential number of terms. Consequently, the time complexity of computing the multinomial NML is dominated by (2.14).

In Paper I, a recursion formula for removing the exponentiality of $\mathcal{C}(\mathcal{M}(K), n)$ was presented. This formula is given by

$$\mathcal{C}(\mathcal{M}(K), n) = \sum_{r_1+r_2=0}^{n} \frac{n!}{r_1! r_2!} \left(\frac{r_1}{n}\right)^{r_1} \left(\frac{r_2}{n}\right)^{r_2} \cdot \mathcal{C}(\mathcal{M}(K^*), r_1) \cdot \mathcal{C}(\mathcal{M}(K - K^*), r_2), \quad (3.1)$$

which holds for all $K^* = 1, \ldots, K-1$. A straightforward algorithm based on this formula was then used to compute $\mathcal{C}(\mathcal{M}(K), n)$ in time $\mathcal{O}\left(n^2 \log K\right)$. See Papers I and V for more details.

In Paper II (see also [31]), the quadratic-time algorithm was improved to $\mathcal{O}\left(n \log n \log K\right)$ by writing (3.1) as a convolution-type sum and then using the Fast Fourier Transform algorithm. However, the relevance of this result is unclear due to severe numerical instability problems it easily produces in practice. See Paper II for more details.

Although the algorithms described above have succeeded in removing the exponentiality of the computation of the multinomial NML, they are still superlinear with respect to n. In Paper III the first linear-time algorithm based on the mathematical technique of generating functions was derived for the problem. The algorithm is based on the following theorem:

Theorem 3.1 *The $\mathcal{C}(\mathcal{M}(K), n)$ terms satisfy the recurrence*

$$\mathcal{C}(\mathcal{M}(K+2), n) = \mathcal{C}(\mathcal{M}(K+1), n) + \frac{n}{K} \cdot \mathcal{C}(\mathcal{M}(K), n). \tag{3.2}$$

Proof. See Paper III. □

It is now straightforward to write a linear-time algorithm for computing the multinomial NML $P_{\mathrm{NML}}(\mathbf{x}^n \mid \mathcal{M}(K))$ based on Theorem 3.1. The process is described in Algorithm 1. The time complexity of the algorithm is

Algorithm 1 The linear-time algorithm for computing $P_{\mathrm{NML}}(\mathbf{x}^n \mid \mathcal{M}(K))$.

1: Count the frequencies $h_1, \ldots, h_K$ from the data $\mathbf{x}^n$
2: Compute the likelihood $P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(K))) = \prod_{k=1}^{K} \left(\frac{h_k}{n}\right)^{h_k}$
3: Set $\mathcal{C}(\mathcal{M}(1), n) = 1$
4: Compute $\mathcal{C}(\mathcal{M}(2), n) = \sum_{r_1+r_2=n} \frac{n!}{r_1! r_2!} \left(\frac{r_1}{n}\right)^{r_1} \left(\frac{r_2}{n}\right)^{r_2}$
5: **for** $k = 1$ to $K-2$ **do**
6: Compute $\mathcal{C}(\mathcal{M}(k+2), n) = \mathcal{C}(\mathcal{M}(k+1), n) + \frac{n}{k} \cdot \mathcal{C}(\mathcal{M}(k), n)$
7: **end for**
8: Output $P_{\mathrm{NML}}(\mathbf{x}^n \mid \mathcal{M}(K)) = \frac{P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(K)))}{\mathcal{C}(\mathcal{M}(K), n)}$

clearly $\mathcal{O}\left(n + K\right)$, which is a major improvement over the previous methods. The algorithm is also very easy to implement and does not suffer from any numerical instability problems. See Paper III for more discussion of the algorithm.

3.1.2 NML Approximations

In the previous section we presented exact NML computation algorithms for multinomial data. The time complexity of the most efficient method was shown to be linear with respect to the size of the data, which can sometimes be too slow for demanding tasks. Consequently, it is important to develop efficient approximations to the multinomial NML. The topic of this section is to present three such methods. The first two of the methods, BIC and Rissanen's asymptotic expansion, are well-known, but the third one, called the Szpankowski approximation, is novel. Since we are able to compute the exact NML, it is also possible to assess the accuracy of the three approximations. This comparison is presented in Section 3.1.3.

In the following, we introduce the three approximations and instantiate them for the multinomial model family. It should be noted that BIC and Rissanen's asymptotic expansion are usually considered as approximations to the stochastic complexity, i.e., the negative logarithm of the NML. To make the formulas easier to interpret, we will adopt this established practice.

Bayesian Information Criterion: The *Bayesian Information Criterion (BIC)* [62], also known as the Schwarz criterion, is the simplest of the three approximations. As the name implies, the BIC has a Bayesian interpretation, but it can also be given a formulation in the MDL setting as showed in [55]. It is derived by expanding the log-likelihood function as a second order Taylor series around the maximum likelihood point $\hat{\boldsymbol{\theta}}$ and then integrating this expansion over the parameter space. This procedure is called the *Laplace's method*. The BIC formula is given by

$$-\log P_{\mathrm{BIC}}(\mathbf{x}^n \mid \mathcal{M}) = -\log P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n), \mathcal{M}) + \frac{d}{2}\log n + \mathcal{O}(1), \quad (3.3)$$

where d is the Euclidean dimension of the parameter space, i.e., the number of parameters. Looking at (3.3), we can see that it contains the same maximum likelihood term as the exact NML equation (2.3). Therefore, the second term $\frac{d}{2}\log(n)$ can be interpreted as an approximation to the parametric complexity.

The instantiation of the BIC approximation for the multinomial case is trivial. If the multinomial variable has K possible values, the number of parameters is $K-1$ and

$$-\log P_{\mathrm{BIC}}(\mathbf{x}^n \mid \mathcal{M}(K)) = -\log P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n), \mathcal{M}(K)) + \frac{K-1}{2}\log n + \mathcal{O}(1). \quad (3.4)$$

The main advantage of BIC is that it is very simple, intuitive and quick to compute. However, it is widely acknowledged that in model selection tasks BIC favors overly simple models (see, e.g., [68]).

Rissanen's Asymptotic Expansion: As proved in [56], for model classes that satisfy certain regularity conditions, a sharper asymptotic expansion than BIC can be derived for the NML. The most important regularity condition is that the Central Limit Theorem should hold for the maximum likelihood estimators for all the elements in the model class. The precise regularity conditions can be found in [56]. Rissanen's asymptotic expansion is given by

$$\begin{aligned} &-\log P_{\mathrm{RIS}}(\mathbf{x}^n \mid \mathcal{M}) = \\ &\quad -\log P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n), \mathcal{M}) + \frac{d}{2}\log\frac{n}{2\,\pi} + \log\int \sqrt{|I(\theta)|}d\theta + o\,(1), \end{aligned} \tag{3.5}$$

where the integral goes over the parameter space Θ. The matrix $I(\theta)$ is called the (expected) *Fisher information matrix* defined by

$$I(\theta) = -E_\theta\left[\frac{\partial^2 \log P(\mathbf{x}^n \mid \theta, \mathcal{M})}{\partial\theta_i\theta_j}\right], \tag{3.6}$$

where θ_i, θ_j go through all the possible pairs of parameters and the expectation is taken over the data space $\mathcal{X}$. The first two terms of (3.5) are essentially the same as in the BIC approximation (3.3). The crucial distinction is the integral term measuring the complexity that comes from the local geometrical properties of the model space. For a more precise discussion of the interpretation of this term, see [21]. Note that unlike the BIC approximation, Rissanen's expansion is *asymptotically correct*. This means that the error in the approximation vanishes as n goes to infinity.

Rissanen's asymptotic expansion for the $\mathcal{M}(K)$ model class was derived in [56], and it is given by

$$\begin{aligned} &-\log P_{\mathrm{RIS}}(\mathbf{x}^n \mid \mathcal{M}(K)) = \\ &-\log P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n), \mathcal{M}(K)) + \frac{K-1}{2}\log\frac{n}{2\pi} + \log\frac{\pi^{K/2}}{\Gamma(K/2)} + o\,(1), \end{aligned} \tag{3.7}$$

where $\Gamma(\cdot)$ is the *Euler gamma function* (see, e.g., [1]). This approximation is clearly very efficient to compute as well. Note that the derivation of the Rissanen's expansion for the Naive Bayes can be found in Paper I.

Szpankowski Approximation: An advanced mathematical tool called *singularity analysis* [16] can be used to derive an arbitrarily accurate ap-

proximation to the multinomial NML. Appendix A.4 gives a brief overview of the method. The Szpankowski approximation is based on a theorem on redundancy rate for memoryless sources [66], which gives

$$\begin{aligned} -\log P_{\mathrm{SZP}}(\mathbf{x}^n \mid \mathcal{M}(K)) = &-\log P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n), \mathcal{M}(K)) \\ &+ \frac{K-1}{2}\log\frac{n}{2} + \log\frac{\sqrt{\pi}}{\Gamma(K/2)} + \frac{\sqrt{2}K\cdot\Gamma(K/2)}{3\Gamma\left(\frac{K}{2}-\frac{1}{2}\right)}\cdot\frac{1}{\sqrt{n}} \\ &+ \left(\frac{3+K(K-2)(2K+1)}{36} - \frac{\Gamma^2(K/2)\cdot K^2}{9\Gamma^2\left(\frac{K}{2}-\frac{1}{2}\right)}\right)\cdot\frac{1}{n} \\ &+ \mathcal{O}\left(\frac{1}{n^{3/2}}\right). \end{aligned} \tag{3.8}$$

The full derivation of this approximation is given in Appendix B. Note that (3.8) is not a general NML approximation. It is only applicable for the multinomial case.

3.1.3 Comparison of the Approximations

As noted in the previous section, the ability to compute the exact NML for the multinomial model gives us a unique opportunity to test how accurate the NML approximations really are. The first thing to notice is that since all the three presented approximations contain the maximum likelihood term, we can ignore it in the comparisons and concentrate on the parametric complexity. Notice that we therefore avoid the problem of trying to choose representative and unbiased data sets for the experiments.

We conducted two sets of experiments with the three approximations. Firstly, we studied the accuracy of the approximations as a function of the size of the data n. In the second set of the experiments we varied the number of values of the multinomial variable. For all the experiments, the following names are used for the three approximations:

- BIC: Bayesian information criteria (3.4)
- RIS: Rissanen's asymptotic expansion (3.7)
- SZP: Szpankowski approximation (3.8)

The results of the first set of experiments can be seen in Figure 3.1, where the difference between the approximative and exact parametric complexity is plotted when the number of values K is set to 2, 4 and 9, respectively. In each figure the size of data n varies from 1 to 100. From the figures we can see that the SZP approximation is clearly the best of the

three. This is naturally anticipated since SZP is theoretically the most accurate one. What might be surprising is the absolute accuracy of SZP. The error is practically zero even for very small values of n. The second best of the approximations is RIS converging monotonically towards the exact value. However, this convergence gets slower when K increases. The figures also nicely show the typical behaviour of the BIC approximation. When the test setting becomes more complex (for $K > 3$), BIC starts to overestimate the parametric complexity.

In the second set of experiments we studied the accuracy of the three approximations when the number of values K varies from 2 to 10. Figure 3.2 shows the difference between the approximative and exact parametric complexity when the size of the data n is fixed to 25, 100 and 500, respectively. Naturally, the accuracy of the SZP approximation is superior in these tests as well. The most dramatic thing to notice from the figures is the rapid decrease in the accuracy of the BIC approximation when K increases. This is in contrast with the RIS approximation, which clearly gets more accurate with increasing amount of data, as anticipated by the theory.

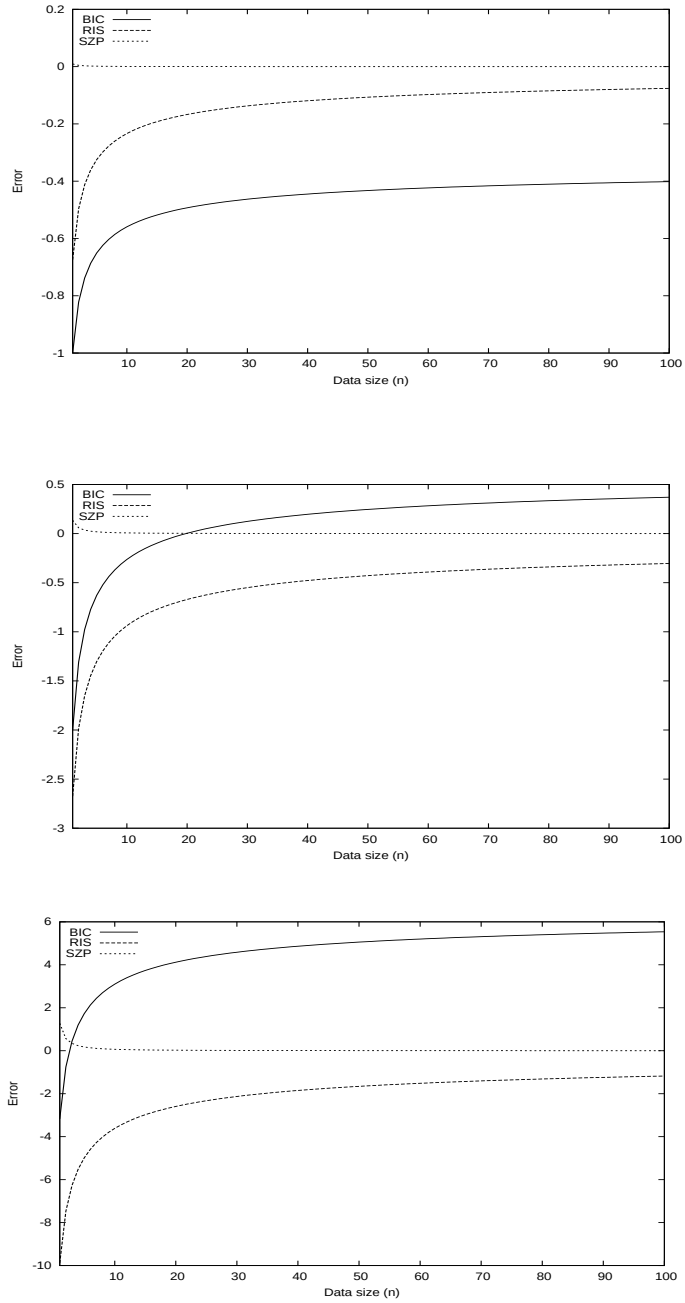


Figure 3.1: The accuracy of the three approximations as a function of the size of the data for $K = 2, 4$ and 9.

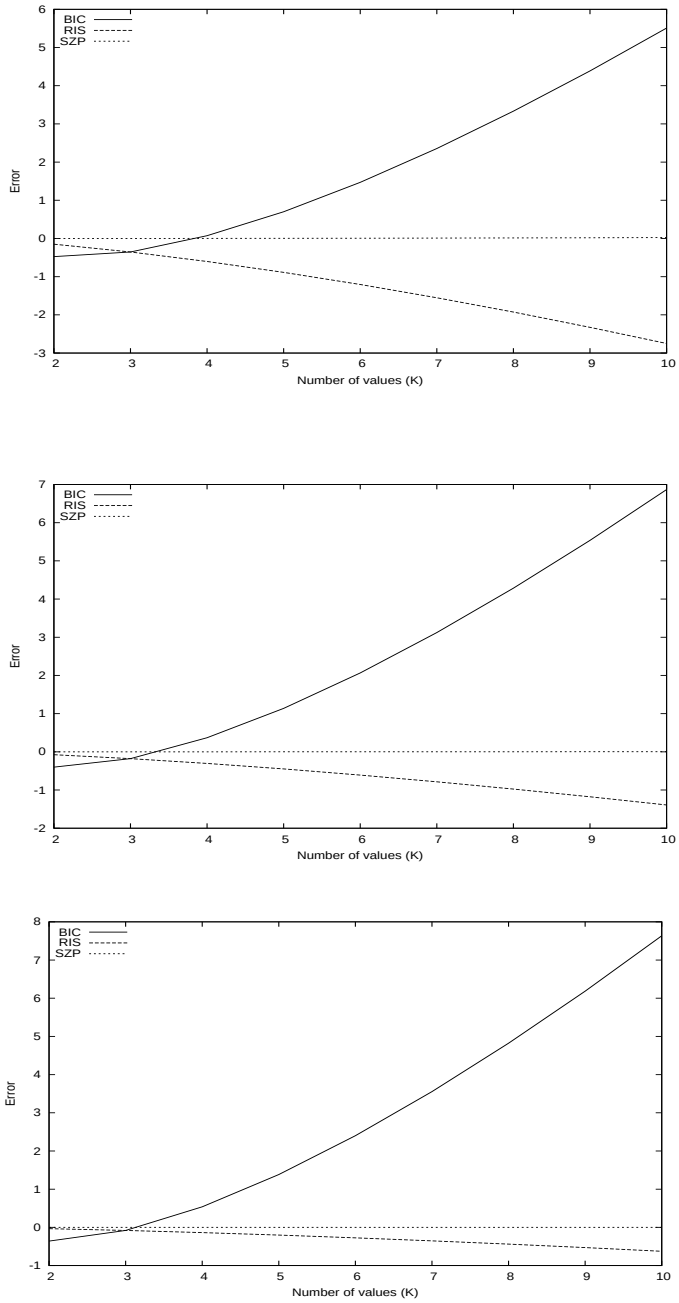


Figure 3.2: The accuracy of the three approximations as a function of the number of values. From top to bottom, the data size n is fixed to 25, 100 and 500.

3.2 The Naive Bayes Model Family

It is clear that the time complexity of computing the NML for the Naive Bayes model family (2.20) is also dominated by the parametric complexity $\mathcal{C}(\mathcal{M}(K_0, K_1, \ldots, K_m), n)$. It turns out (see Papers I and V) that the recursive formula (3.1) can be generalized to this case:

Theorem 3.2 *The terms $\mathcal{C}(\mathcal{M}(K_0, K_1, \ldots, K_m), n)$ satisfy the recurrence*

$$\mathcal{C}(\mathcal{M}(K_0, K_1, \ldots, K_m), n) = \sum_{r_1+r_2=n} \frac{n!}{r_1! r_2!} \left(\frac{r_1}{n}\right)^{r_1} \left(\frac{r_2}{n}\right)^{r_2} \cdot \mathcal{C}_{NB}(\mathcal{M}(K^*, K_1, \ldots, K_m), r_1) \cdot \mathcal{C}_{NB}(\mathcal{M}(K_0 - K^*, K_1, \ldots, K_m), r_2), \tag{3.9}$$

where $K^* = 1, \ldots, K_0 - 1$.

Proof. See Papers I and V. □

In many practical applications of the Naive Bayes the quantity K_0 is unknown. Its value is typically determined as a part of the model class selection process. Consequently, it is necessary to compute NML for model classes $\mathcal{M}(K_0, K_1, \ldots, K_m)$, where K_0 has a range of values, say, $K_0 = 1, \ldots, K_{\max}$. The process of computing NML for this case is described in Algorithm 2. The time complexity of the algorithm is $\mathcal{O}\left(n^2 \cdot K_{\max}\right)$. If the value of K_0 is fixed, the time complexity drops to $\mathcal{O}\left(n^2 \cdot \log K_{\max}\right)$. See Paper V for more details.

Deriving accurate approximations to the Naive Bayes NML is more challenging than in the multinomial case. BIC and the Rissanen's asymptotic expansion can be computed for the Naive Bayes (see Paper I), but the equivalent of the Szpankowski approximation for the multinomial model family (3.8) has not been found. One simple approach is presented in Paper I, where the multinomial NML terms in Algorithm 2 are replaced by the approximations using (3.8). However, the time complexity of the resulting algorithm is still quadratic with respect to the size of the data.

Algorithm 2 The algorithm for computing the NML for the Naive Bayes model family for $K_0 = 1, \ldots, K_{\max}$.

Compute $\mathcal{C}(\mathcal{M}(k), j)$ for $k = 1, \ldots, V_{\max},\ j = 0, \ldots, n$, where $V_{\max} = \max\{K_1, \ldots, K_m\}$
for $K_0 = 1$ to $K_{\max}$ **do**
  Count the frequencies $h_1, \ldots, h_{K_0}, f_{ik1}, \ldots, f_{ikK_i}$ for $i = 1, \ldots, m,\ k = 1, \ldots, K_0$ from the data $\mathbf{x}^n$
  Compute the likelihood: $P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(K_0, K_1, \ldots, K_m)))$ $= \prod_{k=1}^{K_0} \left(\frac{h_k}{n}\right)^{h_k} \prod_{i=1}^{m} \prod_{l=1}^{K_i} \left(\frac{f_{ikl}}{h_k}\right)^{f_{ikl}}$
  Set $\mathcal{C}(\mathcal{M}(K_0, K_1, \ldots, K_m), 0) = 1$
  if $K_0 = 1$ **then**
    Compute $\mathcal{C}(\mathcal{M}(1, K_1, \ldots, K_m), j) = \prod_{i=1}^{m} \mathcal{C}(\mathcal{M}(K_i), j)$ for $j = 1, \ldots, n$
  else
    Compute $\mathcal{C}(\mathcal{M}(K_0, K_1, \ldots, K_m), j)$ $= \sum_{r_1+r_2=j} \frac{j!}{r_1! r_2!} \left(\frac{r_1}{j}\right)^{r_1} \left(\frac{r_2}{j}\right)^{r_2} \cdot \mathcal{C}(\mathcal{M}(1, K_1, \ldots, K_m), r_1)$ $\cdot \mathcal{C}(\mathcal{M}(K_0 - 1, K_1, \ldots, K_m), r_2)$ for $j = 1, \ldots, n$
  end if
  Output $P_{\mathrm{NML}}(\mathbf{x}^n \mid \mathcal{M}(K_0, K_1, \ldots, K_m)) = \frac{P(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n, \mathcal{M}(K_0, K_1, \ldots, K_m)))}{\mathcal{C}(\mathcal{M}(K_0, K_1, \ldots, K_m), n)}$
end for

Chapter 4

MDL Applications

In this chapter, we will show how the NML can be applied to practical problems using the techniques described in Chapter 3. Due to the computational efficiency problems, there are relatively few applications of NML. However, the existing applications have proven that NML works very well in practice and in many cases provides superior results when compared to alternative approaches.

We mention here some examples of NML applications. First, in Papers V and VI, NML was used for clustering of multi-dimensional data and its performance was compared to the Bayesian approach. The results showed that the performance of the NML was especially impressive with small sample sizes. Second, in [60], NML was applied to wavelet denoising of digital images. Since the MDL principle in general can be interpreted as separating information from noise, this approach is very natural. Bioinformatical applications include [43] and [67], where NML was used for DNA sequence compression and data analysis in genomics, respectively. A scheme for using NML for histogram density estimation was presented in Paper IV. In this work, the density estimation problem was regarded as a model class selection task. This approach allowed finding NML-optimal histograms with variable-width bins in a computationally efficient way. Finally, in [12] NML histograms were used for modeling the attributes of the Naive Bayes classifier.

In the following, we will concentrate on two applications: histogram density estimation and clustering of multi-dimensional data. A computationally efficient NML approach for histogram density estimation was proposed in Paper IV. A theoretically interesting recursion formula derived in Paper III was shown to provide a way to compute the NML for histograms in linear time with respect to the sample size. The NML clustering framework was introduced in Paper V. The optimization aspect of the clustering prob-

lem was studied in Paper VI, where five algorithms for efficiently searching the exponentially-sized clustering space were empirically compared.

4.1 Histogram Density Estimation

Density estimation is one of the central problems in statistical inference and machine learning. Given a sample of observations, the goal of *histogram density estimation* is to find a piecewise constant density that describes the data best according to some pre-determined criterion. Although histograms are conceptually simple densities, they are very flexible and can model complex properties like multi-modality with a relatively small number of parameters. Furthermore, one does not need to assume any specific form for the underlying density function: given enough bins, a histogram estimator adapts to any kind of density.

The NML approach for irregular (variable-width bin) histogram density estimation described in Paper IV regards the problem as a model class selection task, where the possible sets of cut points (bin borders) are considered as model classes. The model parameters are the bin masses, or equivalently the bin heights. The NML criterion for comparing candidate histograms can be computed efficiently using the recursion formula derived in Paper III, where the problem of computing the parametric complexity for multinomial model was studied.

There is obviously an exponential number of different cut point sets. Therefore, a brute-force search is not feasible. In Paper IV it was shown that the NML-optimal cut point locations can be found via dynamic programming in a polynomial (quadratic) time with respect to the size of the set containing the cut points considered in the optimization process.

The histogram density estimation is naturally a well-studied problem, but unfortunately almost all of the previous studies, e.g. [6, 23, 73], consider regular (equal-width bin) histograms only. Most similar to our work is [59], in which irregular histograms are learned with the Bayesian mixture criterion using a uniform prior. The same criterion is also used in [23], but the histograms are equal-width only. It should be noted that this difference is significant as the Bayesian mixture criterion does not possess the optimality properties of the NML.

4.1.1 Definitions

Consider a sample of n outcomes $\mathbf{x}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ on the interval $[\mathbf{x}_{\min}, \mathbf{x}_{\max}]$. Without any loss of generality, we assume that the data is sorted into increasing order. Furthermore, we assume that the data is recorded at a

finite accuracy ϵ. This assumption is made to simplify the mathematical formulation, and as can be seen later, the effect of the accuracy parameter ϵ on the stochastic complexity is a constant that can be ignored in the model selection process.

Let $C = (c_1, \ldots, c_{K-1})$ be an increasing sequence of points partitioning the range $[\mathbf{x}_{\min} - \epsilon/2, \mathbf{x}_{\max} + \epsilon/2]$ into the following K intervals (bins):

$$([\mathbf{x}_{\min} - \epsilon/2, c_1],]c_1, c_2], \ldots,]c_{K-1}, \mathbf{x}_{\max} + \epsilon/2]). \tag{4.1}$$

The points c_k are called the *cut points* of the histogram. Define $c_0 = \mathbf{x}_{\min} - \epsilon/2, c_K = \mathbf{x}_{\max} + \epsilon/2$ and let $L_k = c_k - c_{k-1}, k = 1, \ldots, K$ be the bin lengths. Given a parameter vector $\theta \in \Theta$,

$$\Theta = \{(\theta_1, \ldots, \theta_K) : \theta_k \geq 0, \theta_1 + \cdots + \theta_K = 1\}, \tag{4.2}$$

and a set (sequence) of cut points C, we now define the histogram density f_h by

$$f_h(x \mid \theta, C) = \frac{\epsilon \cdot \theta_k}{L_k}, \tag{4.3}$$

where $x \in]c_{k-1}, c_k]$. Note that (4.3) does not define a density in the purest sense, since $f_h(x \mid \theta, C)$ is actually the probability that x falls into the interval $]x - \epsilon/2, x + \epsilon/2]$.

Given (4.3), the likelihood of the whole data sample $\mathbf{x}^n$ is easy to write. We have

$$f_h(\mathbf{x}^n \mid \theta, C) = \prod_{k=1}^{K} \left(\frac{\epsilon \cdot \theta_k}{L_k} \right)^{h_k}, \tag{4.4}$$

where h_k is the number of data points falling into bin k.

4.1.2 NML Histogram

To instantiate the NML distribution (2.3) for the histogram density f_h, we need to find the maximum likelihood parameters $\hat{\boldsymbol{\theta}}(x^n) = (\hat{\theta}_1, \ldots, \hat{\theta}_K)$ and an efficient way to compute the parametric complexity. It is well-known that the ML parameters are given by the relative frequencies $\hat{\theta}_k = h_k/n$, so that we have

$$f_h(\mathbf{x}^n \mid \hat{\boldsymbol{\theta}}(\mathbf{x}^n), C) = \prod_{k=1}^{K} \left(\frac{\epsilon \cdot h_k}{L_k \cdot n} \right)^{h_k}. \tag{4.5}$$

Denote now the parametric complexity of a K-bin histogram by $\log \mathcal{C}(H_K, n)$. We now have the following theorem:

Theorem 4.1 *The term $\mathcal{C}(H_K, n)$ is given by*

$$\mathcal{C}(H_K, n) = \sum_{h_1+\cdots+h_K=n} \frac{n!}{h_1! \cdots h_K!} \prod_{k=1}^{K} \left(\frac{h_k}{n}\right)^{h_k}, \tag{4.6}$$

i.e., the same as the parametric complexity of a K-valued multinomial model.

Proof. See research paper IV. □

This result means that we can compute the parametric complexity for histogram densities using Algorithm 1.

We are now ready to write down the stochastic complexity (2.6) for the histogram model. We have

$$\mathrm{SC}(\mathbf{x}^n \mid C) = -\log \frac{\prod_{k=1}^{K} \left(\frac{\epsilon \cdot h_k}{L_k \cdot n}\right)^{h_k}}{\mathcal{C}(H_K, n)} \tag{4.7}$$

$$= \sum_{k=1}^{K} -h_k(\log(\epsilon \cdot h_k) - \log(L_k \cdot n)) + \log \mathcal{C}(H_K, n). \tag{4.8}$$

Equation (4.8) is the basis for measuring the quality of NML histograms, i.e., comparing different cut point sets. It should be noted that as the term $\sum_{k=1}^{K} -h_k \log \epsilon = -n \log \epsilon$ is a constant with respect to C, the value of ϵ does not affect the comparison.

The histogram density estimation problem is now straightforward to define: find the cut point set C which optimizes the given goodness criterion. In our case the criterion is based on the stochastic complexity (4.8), and the cut point sets are considered as model classes. In practical model class selection tasks, however, the stochastic complexity criterion itself may not be sufficient. The reason is that it is also necessary to encode the model class index in some way, as argued in [21]. We assume that the model class index is encoded with a uniform distribution over all the cut point sets of the same size. For a K-bin histogram with E possible cut points, there are clearly $\binom{E}{K-1}$ ways to choose the cut points. Thus, the codelength for encoding them is $\log \binom{E}{K-1}$.

After these considerations, we define the final criterion (or score) used for comparing different cut point sets as

$$B(\mathbf{x}^n \mid E, K, C) = \mathrm{SC}(\mathbf{x}^n \mid C) + \log \binom{E}{K-1}$$

$$= \sum_{k=1}^{K} -h_k \left(\log(\epsilon \cdot h_k) - \log(L_k \cdot n)\right) + \log \mathcal{C}(H_K, n) + \log \binom{E}{K-1}. \tag{4.9}$$

It is clear that there are an exponential number of possible cut point sets, and thus an exhaustive search to minimize (4.9) is not feasible. However, the optimal cut point set can be found via dynamic programming, which works by tabulating partial solutions to the problem. The final solution is then found recursively. For details, see Paper IV.

To demonstrate the behaviour of the NML histogram method in practice we implemented the dynamic programming algorithm and ran some simulation tests (see Paper IV). We generated data samples of various size from densities of different shapes and then used the dynamic programming method to find the NML-optimal histograms. Figure 4.1 shows two examples of the generating densities (labeled gm6 and gm8) and the corresponding NML-optimal histograms. The sample size is fixed to 10000, and

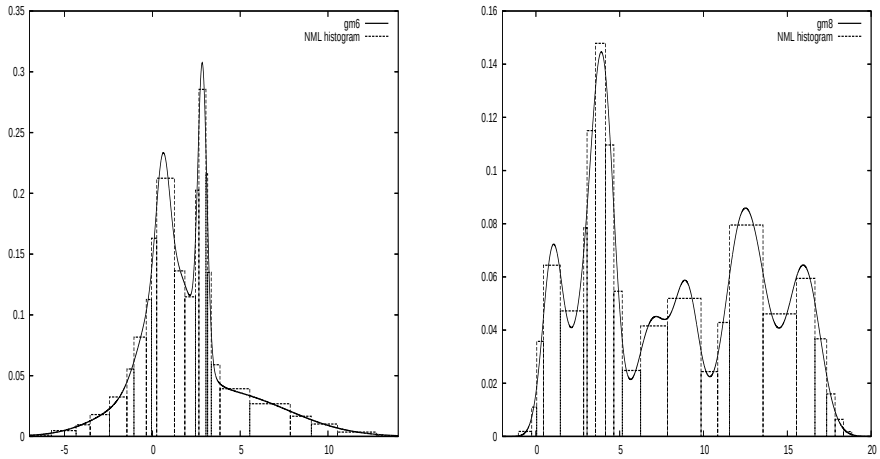


Figure 4.1: The generating densities gm6 and gm8 and the corresponding NML-optimal histograms.

the generating densities are Gaussian finite mixtures with 6 and 8 components, respectively. From the plots we can see that the NML histogram method is able to capture properties such as multi-modality and long tails. Another nice feature is that the algorithm automatically places more bins to the areas where more detail is needed like the high, narrow peaks of gm6. See Paper IV for more empirical tests and discussion.

4.2 Clustering

A *clustering* is a partitional data assignment or data labeling problem, where the goal is to partition the data into mutually exclusive clusters so that similar data vectors are grouped together. The number of clusters is unknown, and determining the optimal number is part of the clustering problem. The data are assumed to be in a vector form so that each data item is a vector consisting of a fixed number of attribute values. Within this framework two fundamental problems can be identified: how to define the goodness of a clustering (data partitioning) and how to find good clusterings with respect to the chosen scoring criterion.

Traditionally, the scoring problem has been approached by first fixing a distance metric, and then by defining a global goodness measure based on this distance metric. However, although this approach is intuitively quite appealing, from the theoretical point of view it introduces many problems such as choosing a suitable distance metric and the handling of non-continuous attributes. A completely different approach to clustering is offered by the *model-based approach*, where for each cluster a data generating function (a probability distribution) is assumed, and the clustering problem is defined as the task to identify these distributions (see, e.g., [64, 18, 7]). In other words, the data are assumed to be generated by a finite mixture model [13, 69, 44]. In this framework the optimality of a clustering can be defined as a function of the fit of data with the finite mixture model, not as a function of the distances between the data vectors.

In Paper V we proposed an NML-based approach for clustering. Intuitively, the idea is that a good clustering is such that one can encode the cluster labels *together* with the data so that the resulting code length is minimized. When the cluster labels are fixed, the finite mixture model is essentially the same as the Naive Bayes model, which allows the use of the techniques described in Section 3.2 for efficient computation of the NML criteria.

The optimization part of the clustering problem, i.e., how to find good clusterings with respect to the NML score, was studied in Paper VI. In that work, five algorithms were proposed to the problem and their performance was compared via empirical tests using several real-world datasets. In Section 4.2.2 we shortly summarize these empirical results.

4.2.1 NML Clustering

Let us assume that our problem domain consists of m discrete variables $X_1, \ldots, X_m$ and that the variable X_i has K_i values. The data $\mathbf{x}^n =$

$(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ consists of observations $\mathbf{x}_j = (x_{j1}, \ldots, x_{jm}) \in \mathcal{X}$, where

$$\mathcal{X} = \{1, 2, \ldots, K_1\} \times \cdots \times \{1, 2, \ldots, K_m\}. \tag{4.10}$$

We assume that the possibly originally continuous variables have been discretized. One reason for focusing on discrete data is that in this case we can model the domain variables by multinomial distributions without having to make restricting assumptions about unimodality, normality etc., which is the situation we face in the continuous case.

A *clustering* of the data set $\mathbf{x}^n$ is defined as a partitioning of the data into mutually exclusive subsets, the union of which forms the data set. The number of subsets is a priori unknown. The *clustering problem* is the task to determine the number of subsets, and to decide to which cluster each data vector belongs.

Formally, we can notate a clustering by using a *clustering vector* $z^n = (z_1, \ldots, z_n)$, where z_j denotes the index of the cluster to which the data vector $\mathbf{x}_j$ is assigned to. Denote the *clustering variable* by X_0 so that z^n is a sample from the distribution of X_0. The number of clusters, say K_0, is implicitly defined in the clustering vector, as it can be determined by counting the number of different values appearing in z^n.

In Paper V, we suggested the following NML-based criterion for finding the optimal clustering $\hat{z}^n$:

$$\hat{z}^n = \arg\max_{z^n} P_{\mathrm{NML}}(\mathbf{x}^n, z^n \mid \mathcal{M}(K_0, K_1, \ldots, K_m)), \tag{4.11}$$

where $\mathcal{M}(K_0, K_1, \ldots, K_m)$ is the Naive Bayes model family with K_0 components. In the clustering framework this means that the data vectors should be partitioned so that the vectors belonging to the same cluster can be compressed well together, i.e., that those data vectors that obey the same set of underlying regularities are grouped together.

Naturally, the criteria for comparing different clusterings can be based on other approaches like Bayesian statistics. In the Bayesian case, the NML distribution in (4.11) is replaced by the Bayesian marginal likelihood (see, e.g. [8, 24]). The approaches were compared empirically in Paper V, where it was shown that NML produces the best results especially with small sample sizes.

4.2.2 Comparison of Clustering Algorithms

The space of potential clusterings is obviously exponential in size, which means that in practice we need to resort to combinatorial search algorithms in our attempt to solve the clustering problem. The search algorithm used

in the empirical tests in Paper V was a simple stochastic greedy algorithm. In Paper VI, we compared five different algorithms for finding good clusterings using several real-world datasets from the UCI repository [2]. Two sets of results were presented. The first set concentrates on finding the number of clusters and the actual clustering minimizing the NML score (4.11). In the second set of experiments, we tested how long it takes for each of the five algorithms to find the respective maximum NML value.

The first search algorithm candidate is a simple stochastic greedy (SG) algorithm suggested in Paper V. Since our definition of clustering is based on the finite mixture model, the standard mixture learning algorithm, EM (Expectation-Maximization) (See [11, 41]) is a natural choice as a second clustering algorithm. The third candidate algorithm is the K-means algorithm (KM), sometimes called the CEM algorithm [45]. It is a simple modification to the EM algorithm.

Each of the algorithms mentioned above needs to be initialized prior to the iterative updating procedure. In our tests, we started each algorithm simply by choosing a random clustering. To test the importance of the initialization, we added two hybrid methods to our set of candidate search algorithms. The first hybrid algorithm (KMSG) starts by running the K-means algorithm until convergence and then switches to the stochastic greedy search. The second algorithm (EMSG) is the same except that the EM algorithm is used as an initializer.

Having fixed the set of candidate search algorithms, the next task is to define a strategy for finding the optimal number of clusters and the actual clustering. Since all the five algorithms converge to a local optimum of the stochastic complexity, the natural strategy is to restart the algorithms several times from different starting points.

Although the NML scoring criterion can be used for comparing clusterings with different number of clusters, the framework does not offer an explicit way to directly infer the optimal number of clusters (K). Consequently, the second part of our search strategy is to vary the parameter K. The complete search strategy is described in Algorithm 3.

In the first batch of results we tested which of the five algorithms find the best clusterings in terms of the stochastic complexity. The results showed that all five candidate algorithms end up choosing a similar number of clusters. However, when we looked at the actual SC values, there were significant differences between the algorithms. Since SC can be interpreted as a quality of a clustering, these differences are important. The hybrid EMSG was clearly the best one of the algorithms, especially with more complex cases, i.e., when the size of data and the optimal number of clusters

Algorithm 3 The search strategy used in our tests.

```
repeat
  for all D in datasets do
    for K = 1 to 20 do
      Choose a random initial K-clustering for dataset D
      for all A in {SG, KM, EM, KMSG, EMSG} do
        Run the algorithm A until converged
      end for
    end for
  end for
until 50 restarts have been made
```

was bigger. Another interesting observation is that the traditional KM and EM algorithms were clearly the worst of the candidate algorithms.

In the second set of experiments we recorded how much CPU time each algorithm required for finding their respective optimal clustering. The most important thing to notice from these results was that the hybrid EMSG algorithm, which in the first batch of empirical results was found to produce comparable or better results than SG, was almost always significantly faster than the SG algorithm proving the intuitive argument that choosing a good initial clustering is important. This made the EMSG algorithm a clear overall winner in our experiments. It is also noteworthy that KM and EM were often much slower than the other algorithms even though they produced inferior results. This makes the applicability of KM and EM even more questionable in the setting used here. See Paper VI for all the details of the empirical tests.

Chapter 5

Conclusion

The Normalized Maximum Likelihood (NML) distribution offers a universal, minimax optimal approach to statistical modeling. In this thesis we have surveyed efficient algorithms for computing the NML in the case of discrete data sets and two model families of practical importance. The first model family we discussed is the multinomial, which can be applied to problems such as density estimation and discretization. In this case, the NML can be computed in linear time. For the Naive Bayes model family, the NML can be computed in quadratic time. Models of this type have been used extensively in clustering or classification domains with good results.

To demonstrate the applicability of the computation algorithms presented, we also discussed two NML applications. The first application was an information-theoretic framework for histogram density estimation. The selected approach based on the MDL principle has several advantages. Firstly, the MDL criterion for model class selection (stochastic complexity) has nice theoretical optimality properties. Secondly, by regarding histogram density estimation as a model class selection problem, it is possible to learn generic, variable-width bin histograms and also estimate the optimal bin count automatically. Furthermore, the MDL criterion itself can be used as a measure of quality of a density estimator, which means that there is no need to assume anything about the underlying generating density. Since the model selection criterion is based on the NML distribution, there is also no need to specify any prior distribution for the parameters.

The second application we described was NML clustering of data. We suggested a framework for this problem based on the idea that a good clustering is such that it allows efficient compression when the data are encoded together with the cluster labels. We also introduced five optimization algorithms for minimizing the stochastic complexity. Using these algorithms, we

conducted an extensive set of experiments with several real-world datasets. In the first part of the tests we recorded the number of clusters chosen and the quality of the actual clusterings found by the algorithms, while the idea of the second batch of tests was to see how much CPU time each algorithm requires for finding the best solution. In the empirical results we found out that all the five algorithms were useful if the goal is to find the NML-optimal number of clusters. However, the quality of the individual clusterings found by the more traditional KM and EM algorithms was questionable. These algorithms were also found to be slow. The most interesting observation was that the novel hybrid EMSG algorithm produced the best results and was also fast.

The methods presented are especially suitable for problems that involve multi-dimensional discrete data sets. Furthermore, unlike the Bayesian methods, information-theoretic approaches such as ours do not require a prior for the model parameters. This is a most important aspect, as constructing a reasonable parameter prior is a notoriously difficult problem, particularly in domains with little background knowledge. All in all, information theory has been found to offer a natural and successful theoretical framework for applications in general.

In the future, our plan is to extend the current work to more complex cases such as general Bayesian networks, which would allow the use of NML in even more involved modeling tasks. Another natural area of future work is to apply the methods of this thesis to other practical tasks involving large discrete databases and compare the results to other approaches, such as those based on Bayesian statistics.

Appendices

Chapter A

Mathematical Background

The purpose of this appendix is to provide the reader with some mathematical techniques that are used in the other parts of the thesis, especially in Appendix B. The topics covered are complex analysis, formal power series, generating functions and asymptotic analysis.

A.1 Review of Complex Analysis

The theory of functions of a complex variable, also called complex analysis for brevity, is one of the most beautiful as well as useful branches of mathematics. It is an essential part of the mathematical background of physicists, mathematicians, engineers and other scientists. From the theoretical viewpoint this is because many mathematical concepts become clarified and unified when examined in the light of complex analysis. From the applied viewpoint the theory is of tremendous value in the solution of problems such as fluid dynamics, heat flow, aerodynamics, electromagnetic theory and many other fields of science and engineering.

For a computer scientist, the importance of complex analysis comes from the fact that the theory can be applied to, e.g., calculation of finite and infinite sums, analyzing algorithms and finding asymptotic behaviour of sequences. In this thesis complex analysis is used for deriving the accurate NML approximation in Appendix B. The purpose of this appendix is to briefly review the most relevant definitions and theorems of complex analysis. For further reading on the subject we recommend the books [51, 74, 65, 26].

A.1.1 The Complex Numbers and the Complex Plane

The set $\mathbb{C}$ of *complex numbers* is introduced to permit solutions to equations like

$$x^2 + 1 = 0, \tag{A.1}$$

that has no solutions in the set $\mathbb{R}$ of real numbers. A complex number has the form $a+bi$, where a and b are real numbers and i is called the imaginary unit and has the property $i^2 = -1$. If $z = a + bi$, a is called the *real part* of z and b is called the *imaginary part* of z. The symbol z, which can stand for any of a set of complex numbers, is called a *complex variable*.

A complex number $z = a+bi$ is uniquely determined by an ordered pair of real numbers (a, b). Because of this correspondence we can associate z with a point (a, b) in coordinate plane. This plane is then called the *complex plane*. The horizontal or x-axis is called the *real axis* and the vertical or y-axis is called the *imaginary axis*. If P is a point in the complex plane corresponding to the complex number $z = a + bi$, then we see from Figure A.1 that

$$a = r\cos\theta, \quad b = r\sin\theta, \tag{A.2}$$

where $r = \sqrt{a^2 + b^2} = |a + bi|$ is called the *modulus* or *absolute value* of z, and θ is called the *argument* of z. It follows that we can write

$$z = a + bi = r(\cos\theta + i\sin\theta), \tag{A.3}$$

which is called the *polar form* of the complex number z.

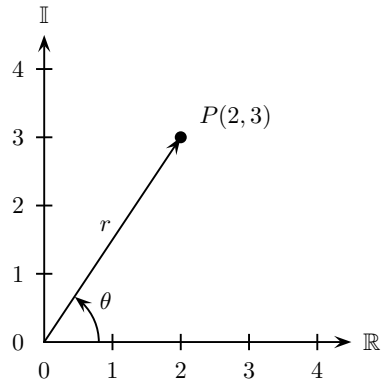


Figure A.1: The polar form of complex number $2 + 3i$.

A.1.2 Roots of Complex Numbers

A number w is called an nth root of a complex number z if $w^n = z$, and we write $w = z^{1/n}$. We can show that if n is a positive integer, then

$$z^{1/n} = (r(\cos\theta + i\sin\theta))^{1/n} \tag{A.4}$$

$$= r^{1/n}\left[\cos\left(\frac{\theta + 2k\pi}{n}\right) + i\sin\left(\frac{\theta + 2k\pi}{n}\right)\right], \tag{A.5}$$

for $k = 0, 1, 2, \ldots, n-1$. It follows that there are n different values for $z^{1/n}$. For example, the five 5th roots of number 32 are

- 2
- $2\left(\cos\frac{2\pi}{5} + i\sin\frac{2\pi}{5}\right)$
- $2\left(\cos\frac{4\pi}{5} + i\sin\frac{4\pi}{5}\right)$
- $2\left(\cos\frac{6\pi}{5} + i\sin\frac{6\pi}{5}\right)$
- $2\left(\cos\frac{8\pi}{5} + i\sin\frac{8\pi}{5}\right)$,

as illustrated in Figure A.2.

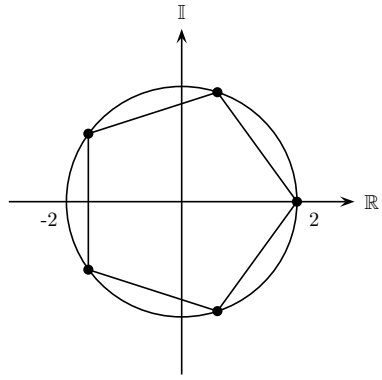


Figure A.2: The 5th roots of complex number 32.

Note that the roots lie on a circle centered at origin of radius $r = 2$ and are spaced at equal angular intervals of $2\pi/5$ radians, i.e., they represent the vertices of a regular pentagon.

A.1.3 Analytic Functions

A *complex function* is a function f whose domain and range are subsets of the set $\mathbb{C}$ of complex numbers. Because $\mathbb{R}$ is a subset of the set $\mathbb{C}$, every real-valued function of a real variable is also a complex function. Furthermore, every complex function can be defined in terms of two real functions $u(a, b)$ and $v(a, b)$ as $f(z) = u(a, b)+iv(a, b)$. This implies that the study of complex functions is closely related to the study of real multivariate functions of two real variables.

Suppose that a complex function f is defined in a deleted neighborhood of a point z_0 and that l is a complex number. The *limit* of f as z tends to z_0 exists and is equal to l, written as $\lim_{z\to z_0} f(z) = l$, if for every $\epsilon > 0$ there exists a number δ such that $|f(z) - l| < \epsilon$ whenever $|z - z_0| < \delta$. Complex and real limits have many common properties, but there is at least one very important difference. For limits of complex functions, z is allowed to approach z_0 from any direction in the complex plane, that is, along any curve or path through z_0. In order that $\lim_{z\to z_0} f(z) = l$, it is required that $f(z)$ approaches the same complex number l along every possible curve through z_0.

The complex derivative is defined similarly as its real counterpart. Suppose that a complex function f is defined in a neighborhood of a point z_0. The *derivative* of f at z_0 is

$$f'(z_0) = \lim_{z\to z_0} \frac{f(z) - f(z_0)}{z - z_0}, \tag{A.6}$$

provided that the limit exists. Furthermore, the function f is said to be *analytic* at a point z_0 if the derivative $f'(z_0)$ exists at z_0 and at every point in some neighborhood of z_0. If f is analytic at every point in an open connected set (domain) D we say that $f(z)$ is analytic in D. The term *holomorphic* is often used as a synonym for analytic. A function that is analytic at every point in the complex plane is said to be an *entire function*.

A remarkable property of analytic functions is the *infinite differentiability*: if f is analytic in a domain D, then f has derivatives of *all* orders in D. This is not necessarily true for functions of real variables. Furthermore, if z_0 is a point in D, then by the *Taylor's theorem*, f has the series representation

$$f(z) = \sum_{n\geq 0} \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n \tag{A.7}$$

valid for the largest circle C with center at z_0 and radius R that lies entirely within D. The number R is called the *radius of convergence*.

A.1.4 Complex Integration

A complex integral is defined in a manner that is quite similar to that of a line integral in the Cartesian plane. Let f be a complex function defined at all points on a smooth curve C. Subdivide C into n parts by means of $z_1, \ldots, z_{n-1}$ chosen arbitrarily. On each arc joining z_{k-1} and z_k choose a point α_k and form a sum

$$S_n = f(\alpha_1)(z_1 - z_0) + f(\alpha_2)(z_2 - z_1) + \cdots + f(\alpha_n)(z_n - z_{n-1}), \quad \text{(A.8)}$$

where z_0 and z_n are the starting and end poinds of C, respectively. On writing $\Delta z_k = z_k - z_{k-1}$, this becomes

$$S_n = \sum_{k=1}^{n} f(\alpha_k)\Delta z_k. \quad \text{(A.9)}$$

Let the number of subdivisions n increase in such a way that the largest of the arc lengths $|\Delta z_k|$ approaches zero. If the sum S_n approaches a limit which does not depend on the choice of the z_k's we call this limit a *complex (line) integral* of f along curve C and denote it by

$$\oint_C f(z)dz. \quad \text{(A.10)}$$

Function f is said to be *integrable* along curve C. If f is analytic at all points of a domain D and if curve C is lying in D then f is certainly integrable along C.

Another remarkable result of complex analysis is the *Cauchy's integral theorem*: Suppose that a function f is analytic at all points within and on a simple closed curve C. Then,

$$\oint_C f(z)dz = 0. \quad \text{(A.11)}$$

A.1.5 Laurent Expansion

If a complex function f fails to be analytic at a point z_0, then this point is said to be a *singularity* of the function f. The Taylor expansion (A.7) does not hold at a singularity point. However, if the singularity z_0 is *isolated*, i.e., there exists some deleted neighborhood of z_0 throughout which f is analytic, it is possible to represent f by a series involving both negative and non-negative integer powers of $z - z_0$. This series is called the *Laurent expansion*,

$$f(z) = \sum_{n=-\infty}^{\infty} a_n(z - z_0)^n. \quad \text{(A.12)}$$

Furthermore, the coefficients a_n are given by

$$a_n = \frac{1}{2\pi i} \oint_C \frac{f(z)dz}{(z - z_0)^{n+1}}, \tag{A.13}$$

where C is any simple closed curve that encloses z_0 and that lies entirely inside a region in which f is analytic.

An isolated singularity z_0 of a complex function f is given a classification depending on whether its Laurent expansion (A.12) contains zero, a finite number, or an infinite number of terms of negative powers.

1. If all the coefficients a_{-n} are zero, then z_0 is called a *removable singularity.*

2. If a finite number, say k, of coefficients a_{-n} are non-zero, then z_0 is called a *pole of order* k.

3. If an infinite number of coefficients a_{-n} are non-zero, then z_0 is called an *essential singularity.*

If the denominator of a rational function f has a zero of order k at z_0, then the function f has a pole of order k at z_0.

A.1.6 The Residue Theorem

The coefficient a_{-1} in the Laurent series (A.12) has a special meaning. This coefficient is called the *residue* of function f at the isolated singularity z_0 and denoted by

$$a_{-1} = \operatorname*{Res}_{z=z_0} f(z). \tag{A.14}$$

The reason why the residue concept is important is that under some circumstances we can evaluate complex integrals by summing the residues at the isolated singularities of a function. More precisely, the *Residue theorem* states that if f is analytic inside and on a simple closed curve C, except at a finite number of isolated singularities $z_1, z_2, \ldots, z_n$ within C, then

$$\oint_C f(z)dz = 2\pi i \sum_{k=1}^{n} \operatorname*{Res}_{z=z_k} f(z). \tag{A.15}$$

Note that the residue theorem is an extension of the Cauchy's integral theorem (A.11).

The residue theory has many applications. It can be used, e.g., to evaluate *real* integrals, to find the locations of zeros of an analytic function,

to sum infinite series and to find integral transforms such as the Laplace transform and its inverse.

There are several ways to calculate residues. Obviously, if we can somehow find the Laurent expansion of a function f at point z_0, we can just pick the coefficient a_{-1} from the series. Otherwise, if the singularity z_0 is a pole of order k, then

$$\operatorname*{Res}_{z=z_0} f(z) = \frac{1}{(k-1)!} \lim_{z \to z_0} \frac{d^{k-1}}{dz^{k-1}} [(z - z_0)^k f(z)]. \tag{A.16}$$

Interestingly, this means that in some cases complex integrals can be evaluated by taking derivatives of complex functions.

A.1.7 Puiseux Expansion

We finalize the discussion on complex analysis by a very special topic of fractional power or *Puiseux* series. This series is relevant in the derivation of the accurate NML approximation in Appendix B. Suppose f is a multivalued analytic function and z_0 its special singularity called *branch point* of order $k-1$. The exact definition of a branch point is complicated and omitted here, but as an example the function $(z-1)^{1/3}$ has a branch point of order 2 at $z_0 = 1$, and the function $\sqrt{z(z-1)}$ has two branch points at 0 and 1, each of order 1. In the neighborhood of a branch point z_0, the function f can be represented as a series

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z - z_0)^{n/k}. \tag{A.17}$$

Note that the series (A.17) is an extension of the Laurent expansion (A.12).

Unfortunately, there is no simple formula for calculating the coefficients of a Puiseux series. For the purposes of this thesis, however, a special result on inversion of Puiseux series presented in [14] is suitable. In that work, series expansions are classified into four types of systematic patterns. We omit the full categorization here, but the category relevant to the main part of the thesis is called "Type II" and it is of form

$$f(z) = a_0 + \sum_{n \geq 1} a_n (z - z_0)^{n-1+\beta}, \tag{A.18}$$

where $\beta > 0$. According to the theorem, the inverse function of f can then be represented as a Puiseux series

$$F(w) = \sum_{n \geq 0} b_n (w - w_0)^{n/\beta}, \tag{A.19}$$

for some sequence of coefficients b_n. Note that $f(z_0) = a_0 = w_0$ and $F(w_0) = z_0$. An example of using the inversion is given in Appendix B.

A.2 Formal Power Series

In this appendix we give a short overview of the theory of formal power series. We concentrate on the issues that are relevant to the other parts of the thesis. Readers interested to learn more about formal power series can refer to, e.g., [71, 19].

A.2.1 Definition

A *formal power series* is an expression of the form

$$\sum_{n\geq 0} a_n z^n, \tag{A.20}$$

where the numbers a_n are called the coefficients of the series. In the theory of formal power series, the variable z is considered as a formal symbol, and the convergence of series (A.20) is not an issue. If, however, the series converges for some values of z, it is a big advantage. For example, the singularity analysis discussed in Appendix A.4 is based on this *analytic theory* of power series. In practice, however, all the operations on series can be performed without worrying about the convergence.

A.2.2 Linear Combination

The most basic of formal power series operations is taking a linear combination of two series. Since formal power series are just infinite polynomials, we have

$$\alpha \sum_{n\geq 0} a_n z^n + \beta \sum_{n\geq 0} b_n z^n = \sum_{n\geq 0} (\alpha a_n + \beta b_n) z^n, \tag{A.21}$$

for numbers α, β.

A.2.3 Multiplication

Another basic operation is multiplication of two or more power series. By basic arithmetics,

$$\left(\sum_{n\geq 0} a_n z^n\right) \cdot \left(\sum_{n\geq 0} b_n z^n\right) = (a_0 + a_1 z + a_2 z^2 + \cdots)(b_0 + b_1 z + b_2 z^2 + \cdots) \tag{A.22}$$

$$= (a_0 b_0) + (a_0 b_1 + a_1 b_0) z \tag{A.23}$$
$$+ (a_0 b_2 + a_1 b_1 + a_2 b_0) z^2 + \cdots$$

$$= \sum_{n\geq 0} \left(\sum_{k=0}^{n} a_k b_{n-k}\right) z^n. \tag{A.24}$$

The series (A.24) is called the *Cauchy product* or *convolution*.

The multiplication operation also generalizes to a product of three or more series. For example, the product of three formal power series is

$$\left(\sum_{n\geq 0} a_n z^n\right) \cdot \left(\sum_{n\geq 0} b_n z^n\right) \cdot \left(\sum_{n\geq 0} c_n z^n\right) \tag{A.25}$$

$$= (a_0 + a_1 z + a_2 z^2 + \cdots)(b_0 + b_1 z + b_2 z^2 + \cdots)(c_0 + c_1 z + c_2 z^2 + \cdots) \tag{A.26}$$

$$= (a_0 b_0 c_0) + (a_0 b_0 c_1 + a_0 b_1 c_0 + a_1 b_0 c_0) z \tag{A.27}$$
$$+ (a_0 b_0 c_2 + a_0 b_1 c_1 + a_0 b_2 c_0 + a_1 b_0 c_1 + a_1 b_1 c_0 + a_2 b_0 c_0) z^2 + \cdots \tag{A.28}$$

$$= \sum_{n\geq 0} \left(\sum_{r+s+t=n} a_r b_s c_t\right) z^n. \tag{A.29}$$

A.2.4 Reciprocal Series

A more complex operation is taking the reciprocal of a formal power series. It is defined as

$$\sum_{n\geq 0} b_n z^n = \frac{1}{\sum_{n\geq 0} a_n z^n}, \tag{A.30}$$

from which it follows that

$$(a_0 + a_1 z + a_2 z^2 + \cdots)(b_0 + b_1 z + b_2 z^2 + \cdots) \equiv 1, \tag{A.31}$$

i.e., the trivial sequence (1, 0, 0, ...). Using the product rule (A.24) we can solve the reciprocal coefficients b_n as

$$a_0 b_0 = 1, \quad b_0 = \frac{1}{a_0} \tag{A.32}$$

$$a_0 b_1 + a_1 b_0 = 0, \quad b_1 = -\frac{a_1 b_0}{a_0} = -\frac{a_1}{a_0^2} \tag{A.33}$$

$$a_0 b_2 + a_1 b_1 + a_2 b_0 = 0, \quad b_2 = -\frac{a_1 b_1 + a_2 b_0}{a_0} = \frac{a_1^2}{a_0^3} - \frac{a_2}{a_0^2}, \tag{A.34}$$

and so on. This result is used in Appendix B. It is easy to see that the reciprocal of a series is only defined when a_0, the constant term in the original series, is non-zero.

As a simple example, we show that the reciprocal of $(1, -1, 0, 0, \ldots)$ is the sequence $(1, 1, 1, \ldots)$, i.e.,

$$\frac{1}{1-z} = \sum_{n \geq 0} z^n. \tag{A.35}$$

This is easy to prove, since

$$(1-z)(1+z+z^2+\cdots) = (1+z+z^2+\cdots) + (-z-z^2+\cdots) \equiv 1. \tag{A.36}$$

A.2.5 Inverse Series

The reciprocal operation is not to be confused with the subtler operation of inverting a series. Inverse of a series

$$f(z) = \sum_{n \geq 0} a_n z^n \tag{A.37}$$

is defined as a series

$$g(z) = \sum_{n \geq 0} b_n z^n, \tag{A.38}$$

if

$$f(g(z)) = g(f(z)) \tag{A.39}$$

$$\begin{aligned} &= a_0 + a_1(b_0 + b_1 z + b_2 z^2 + \cdots) \\ &\qquad + a_2(b_0 + b_1 z + b_2 z^2 + \cdots)^2 + \cdots \equiv z, \end{aligned} \tag{A.40}$$

i.e., the trivial sequence (0, 1, 0, 0, ...). As argued in [71] (Chapter 2.1), this operation only makes sense if the constant terms a_0, b_0 are zero or if f

is a polynomial (finite). Otherwise, the process of finding the coefficients of the inverse series is infinite. Consequently, we have

$$f(g(z)) = a_1(b_1z + b_2z^2 + b_3z^3 + \cdots) + a_2(b_1z + b_2z^2 + b_3z^3 + \cdots)^2 \tag{A.41}$$

$$+ a_3(b_1z + b_2z^2 + b_3z^3 + \cdots)^3 + \cdots$$

$$= (a_1b_1)z + (a_1b_2 + a_2b_1^2)z^2 + (a_1b_3 + 2a_2b_1b_2 + a_3b_1^3)z^3 + \cdots \equiv z, \tag{A.42}$$

from which we get by coefficient comparison

$$a_1b_1 = 1, \quad b_1 = \frac{1}{a_1} \tag{A.43}$$

$$a_1b_2 + a_2b_1^2 = 0, \quad b_2 = -\frac{a_2b_1^2}{a_1} = -\frac{a_2}{a_1^3} \tag{A.44}$$

$$a_1b_3 + 2a_2b_1b_2 + a_3b_1^3 = 0, \quad b_3 = -\frac{2a_2b_1b_2 + a_3b_1^3}{a_1} = \frac{2a_2^2}{a_1^5} - \frac{a_3}{a_1^4}. \tag{A.45}$$

This result is also used in Appendix B.

A.3 Generating Functions

One of the most powerful ways to analyze a sequence of numbers is to form a power series with the elements of the sequence as coefficients. The resulting function is called the *generating function* of the sequence. Generating functions can be seen as a bridge between discrete mathematics and continuous analysis. They can be used for finding recurrence formulas and asymptotic expansions, proving combinatorial identities and finding statistical properties of a sequence.

In this appendix we will present a short overview of generating functions and illustrate their use with several examples. Good sources for further reading on generating functions are [71, 3, 19, 27, 28, 29].

A.3.1 Definition

The (ordinary) generating function of a sequence

$$\langle a_n \rangle = (a_0, a_1, a_2, \ldots) \tag{A.46}$$

is defined as a series

$$A(z) = \sum_{n \geq 0} a_n z^n, \tag{A.47}$$

where z is a dummy symbol (or a complex variable). The importance of generating functions is that the function $A(z)$ is a representation of the whole sequence $\langle a_n \rangle$. By studying this function we can get important information about the sequence, such as asymptotic form of the coefficients.

The most basic generating function is the one generating the constant sequence $(1, 1, 1, \ldots)$. As already shown in Appendix A.2, this function is given by

$$\frac{1}{1-z} = \sum_{n \geq 0} z^n. \tag{A.48}$$

A.3.2 Fibonacci Numbers

As a first non-trivial example of the power of generating functions we consider the famous Fibonacci sequence

$$\langle F_n \rangle = (0, 1, 1, 2, 3, 5, 8, \ldots), \tag{A.49}$$

defined by the recurrence relation

$$F_{n+1} = F_n + F_{n-1}, \qquad (n \geq 1, F_0 = 0, F_1 = 1). \tag{A.50}$$

To find the generating function

$$F(z) = \sum_{n \geq 0} F_n z^n = z + z^2 + 2z^3 + 3z^4 + 5z^5 + 8z^6 + \cdots, \tag{A.51}$$

we multiply the recurrence (A.50) by z^n and sum over $n \geq 1$:

$$\sum_{n \geq 1} F_{n+1} z^n = \sum_{n \geq 1} F_n z^n + \sum_{n \geq 1} F_{n-1} z^n \tag{A.52}$$

$$\frac{F(z) - z}{z} = F(z) + zF(z) \tag{A.53}$$

$$F(z) = \frac{z}{1 - z - z^2}. \tag{A.54}$$

From the basic complex analysis we know that the function $F(z)$ has a *partial fraction expansion* of the form

$$\frac{A}{1 - \alpha z} + \frac{B}{1 - \beta z} = \frac{z}{1 - z - z^2} \tag{A.55}$$

for some numbers α, β, A, B. To find these constants, we write (A.55) as

$$\frac{A}{1 - \alpha z} + \frac{B}{1 - \beta z} = \frac{A(1 - \beta z) + B(1 - \alpha z)}{(1 - \alpha z)(1 - \beta z)} = \frac{z}{1 - z - z^2}. \tag{A.56}$$

For this to hold, we must have

$$(A+B) - (A\beta + B\alpha)z = z \tag{A.57}$$

$$(1-\alpha z)(1-\beta z) = 1 - z - z^2, \tag{A.58}$$

which can be solved straightforwardly as

$$\alpha = \frac{1+\sqrt{5}}{2},\ \beta = \frac{1-\sqrt{5}}{2},\ A = \frac{1}{\sqrt{5}},\ B = -\frac{1}{\sqrt{5}}. \tag{A.59}$$

We can now write

$$F(z) = \frac{A}{1-\alpha z} + \frac{B}{1-\beta z} \tag{A.60}$$

$$= A\sum_{n\geq 0}(\alpha z)^n + B\sum_{n\geq 0}(\beta z)^n \tag{A.61}$$

$$= \sum_{n\geq 0}\left(A\alpha^n + B\beta^n\right) z^n, \tag{A.62}$$

and by plugging the solved values (A.59) into Equation (A.62), we get the closed form solution for the nth Fibonacci number

$$F_n = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right). \tag{A.63}$$

A.3.3 Integer Partitions

Let $q_K(n)$ be the number of partitions of integer n into K parts, i.e., the number of finite non-increasing sequences of non-negative integers $(h_1, \ldots, h_K)$ such that $h_1 + h_2 + \cdots + h_K = n$. For example, $q_3(5) = 5$, since we have

$$5 = 5+0+0 = 4+1+0 = 3+2+0 = 3+1+1 = 2+2+1. \tag{A.64}$$

In this section we want to find the generating function of the numbers $q_K(n)$, i.e.,

$$Q_K(z) = \sum_{n\geq 0} q_K(n) z^n. \tag{A.65}$$

Note that an asymptotic analysis of $Q_K(n)$ is discussed in Appendix A.4.

It is well-known (see, e.g., [3]) that the function generating the numbers $q_K(n)$ is given by

$$Q_K(z) = \frac{1}{1-z}\cdot\frac{1}{1-z^2}\cdot\frac{1}{1-z^3}\cdots\frac{1}{1-z^K}. \tag{A.66}$$

Partition	Star diagram	Conjugate	1.term	2.term	3.term
$5,0,0$	$\star\ \star\ \star\ \star\ \star$	$1,1,1,1,1$	z^5	1	1
$4,1,0$	$\star\ \star\ \star\ \star$ $\star$	$2,1,1,1$	z^3	z^2	1
$3,2,0$	$\star\ \star\ \star$ $\star\ \star$	$2,2,1$	z	z^4	1
$3,1,1$	$\star\ \star\ \star$ $\star$ $\star$	$3,1,1$	z^2	1	z^3
$2,2,1$	$\star\ \star$ $\star\ \star$ $\star$	$3,2$	1	z^2	z^3

Table A.1: Partitions and conjugate partitions of integer 5 into 3 parts.

Intuitively, this result can be understood via an example. Take the above-mentioned case with $n = 5, K = 3$. The generating function is

$$Q_3(z) = \frac{1}{1-z} \cdot \frac{1}{1-z^2} \cdot \frac{1}{1-z^3} \tag{A.67}$$

$$= (1 + z + z^2 + z^3 + \cdots)(1 + z^2 + z^4 + z^6 + \cdots) \cdot (1 + z^3 + z^6 + z^9 + \cdots). \tag{A.68}$$

By the basic definition of generating functions, it is clear that the coefficient of z^5 in the expansion of (A.68) must be $q_3(5) = 5$. To see that this is indeed the case, take a look at Table A.1, where the partitions of 5 into 3 parts are listed. Each partition of n can be represented as a *star diagram* composed of n stars arranged in rows. The number of stars in each row is determined by the elements of the partition. Counting the stars by columns instead of rows, we get the *conjugate partition* of the original partition. Now, each conjugate partition represents a way to get the term z^5 in (A.68). Take, for example, the conjugate partition $(2,1,1,1)$:

1. The number of 1's in the partition is 3, so pick the 3rd order term from $(1 + z + z^2 + z^3 + \cdots)$, i.e., z^3.

2. The number of 2's in the partition is 1, so pick the 1st order term from $(1 + z^2 + z^4 + z^6 + \cdots)$, i.e., z^2.

3. The number of 3's in the partition is 0, so pick the 0th order term from $(1 + z^3 + z^6 + z^9 + \cdots)$, i.e., 1.

We end up with the term $z^3 \cdot z^2 \cdot 1 = z^5$, as desired. The other four partitions are treated similarly. We can therefore conclude that the function $Q_3(z)$ generates numbers $q_3(n)$. The full proof can be found in [3].

A.4 Asymptotic Analysis of Generating Functions

In this appendix we will present methods for finding asymptotic behaviour of a sequence based on the theory of generating functions. For the purposes of this thesis, a powerful method called *singularity analysis* by Flajolet and Odlyzko [16] is especially suitable. Additional sources of information on singularity analysis are [52, 17, 66]. Other asymptotic methods, such as bootstrapping, Tauberian theorems, Darboux's method and the saddle point method are discussed in [10, 20, 19, 71, 66].

Suppose we have found the generating function for a certain sequence of numbers that interests us. The goal of asymptotic analysis is to find a simple function of n which approximates well the values of the sequence when n is large. This can be achieved by analyzing the singularities of the generating function. Suitable asymptotic analysis method is then chosen based on the nature of the singularities.

Especially important is the singularity that is nearest to the origo. As argued in [66], this *dominant singularity* determines the asymptotic growth of the coefficients of the generating function. Therefore, it is only necessary to locate this singularity and analyze the behaviour of the function around it.

A.4.1 Rational Functions

We start the discussion on asymptotic analysis by a relatively simple case of rational generating functions, whose only singularities are poles. Let $f(z)$ be a rational function generating the sequence $\langle a_n \rangle$. Suppose $f(z)$ is analytic at zero and has poles at points $p_1, p_2, \ldots, p_m$. Then there exists m polynomials $(P_1, \ldots, P_m)$ such that exactly

$$a_n = [z^n] f(z) = \sum_{j=1}^{m} P_j(n) p_i^{-n}. \tag{A.69}$$

Furthermore, the degree of P_j is equal to the order of the pole at p_j minus one. In particular, a single pole only contributes a constant term to (A.69). This theorem is proved in, e.g., [66]. In practice, the polynomials P_j can be found via residue calculus.

To illustrate the use of (A.69), let us consider a version of the classic money changing problem: in how many ways can one pay an amount of n cents using only coins of 1, 2 and 5 cents? Let m_n denote this number. To solve the problem, we need to find the generating function $m(z)$ for the sequence $\langle m_n \rangle$. The money changing problem is closely related to the counting of integer partitions discussed in Appendix A.3. Using similar arguments, it is easy to see that the generating function is given by

$$m(z) = \frac{1}{(1-z)(1-z^2)(1-z^5)}, \tag{A.70}$$

which is a rational function and analytic at zero, so (A.69) applies.

The first step is to find the poles of (A.70). From the complex root discussion of Appendix A.1, we have:

- The only pole of $(1-z)$ is 1.
- The poles of $(1-z^2)$ are $1, -1$.
- The poles of $(1-z^5)$ are:
  - 1
  - $\cos\frac{2\pi}{5} + i\sin\frac{2\pi}{5}$
  - $\cos\frac{4\pi}{5} + i\sin\frac{4\pi}{5}$
  - $\cos\frac{6\pi}{5} + i\sin\frac{6\pi}{5}$
  - $\cos\frac{8\pi}{5} + i\sin\frac{8\pi}{5}$.

Thus, the function $m(z)$ has a triple pole at $z = 1$ and several single poles. We choose here to ignore the single poles, since they only contribute a constant term to (A.69).

By the Laurent's theorem presented in Appendix A.1, we know that $m(z)$ has a Laurent expansion at $z = 1$,

$$m(z) = \frac{a_{-3}}{(z-1)^3} + \frac{a_{-2}}{(z-1)^2} + \frac{a_{-1}}{(z-1)} + \sum_{n\geq 0} a_n (z-1)^n. \tag{A.71}$$

The coefficients a_n can be found via basic residue calculus. By the coefficient formula (A.13),

$$a_{-3} = \frac{1}{2\pi i} \oint_C \frac{(z-1)^2}{(1-z)(1-z^2)(1-z^5)} dz \tag{A.72}$$

$$= \frac{1}{2\pi i} \oint_C \frac{1}{(1+z)(1-z)(1+z+z^2+z^3+z^4)} dz, \tag{A.73}$$

which is by the residue theorem (A.15)

$$a_{-3} = \operatorname*{Res}_{z=1} \frac{1}{(1+z)(1-z)(1+z+z^2+z^3+z^4)} \tag{A.74}$$

$$= \lim_{z\to 1} \frac{z-1}{(1+z)(1-z)(1+z+z^2+z^3+z^4)} \tag{A.75}$$

$$= \lim_{z\to 1} \frac{-1}{(1+z)(1+z+z^2+z^3+z^4)} \tag{A.76}$$

$$= -\frac{1}{10}. \tag{A.77}$$

Similarly we can calculate that $a_{-2} = 1/4$ and $a_{-1} = -13/40$. The Laurent expansion is then

$$m(z) = -\frac{1}{10(z-1)^3} + \frac{1}{4(z-1)^2} - \frac{13}{40(z-1)} + \sum_{n\geq 0} a_n (z-1)^n \tag{A.78}$$

$$= \frac{1}{10(1-z)^3} + \frac{1}{4(1-z)^2} + \frac{13}{40(1-z)} + \sum_{n\geq 0} a_n (z-1)^n. \tag{A.79}$$

To extract the nth coefficient from the expansion (A.79), we need the following basic combinatoric result (see, e.g., [71])

$$[z^n] \frac{1}{(1-z)^{k+1}} = \binom{n+k}{n}, \tag{A.80}$$

so (see also Table A.2)

$$[z^n] \frac{1}{(1-z)^3} = \binom{n+2}{n} = \frac{1}{2}n^2 + \frac{3}{2}n + 1, \tag{A.81}$$

$$[z^n] \frac{1}{(1-z)^2} = \binom{n+1}{n} = n + 1. \tag{A.82}$$

Now we get the asymptotics for the money changing problem,

$$m_n \sim \frac{1}{10}\left(\frac{1}{2}n^2 + \frac{3}{2}n + 1\right) + \frac{1}{4}(n+1) + \mathcal{O}(1) \tag{A.83}$$

$$= \frac{1}{20}n^2 + \frac{8}{20}n + \mathcal{O}(1). \tag{A.84}$$

To assess the accuracy of the approximation (A.84), we used Maple to calculate the full expansion of the generating function (A.70) therefore obtaining the exact sequence $\langle m_n \rangle$. The comparison of the exact and asymptotic values is given in Figures A.3 and A.4. Clearly, the approximation works very well.

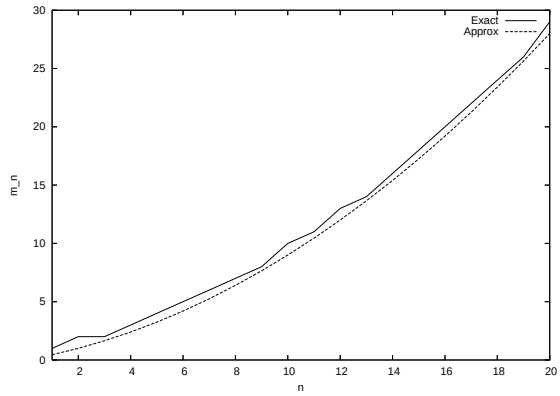


Figure A.3: The comparison of the exact and approximative solutions for the money changing problem with $n = 1, \ldots, 20$.

A.4.2 Asymptotics of Integer Partitions

In this section we briefly discuss the asymptotic analysis of the integer partition generating function (A.66) introduced in Appendix A.3,

$$Q_K(z) = \frac{1}{1-z} \cdot \frac{1}{1-z^2} \cdot \frac{1}{1-z^3} \cdots \frac{1}{1-z^K}. \tag{A.85}$$

Clearly, this function has a pole of order K at $z = 1$. From the discussion of the previous section we know that the highest order pole dominates the asymptotics of rational generating functions. Furthermore, by Equation (A.69) a pole of order K contributes a term of degree $K - 1$. Thus, we can conclude that the number of partitions of an integer n into K parts is $\mathcal{O}\left(n^{K-1}\right)$, i.e., asymptotically the same as the number of compositions.

A.4.3 Algebraic-Logarithmic Functions: The Singularity Analysis

A very general and powerful asymptotic method called *singularity analysis* was introduced in [16]. In its most general form it allows to find asymptotics for *algebraic-logarithmic* functions of the form

$$(1-z)^{-\alpha} \left(\frac{1}{z} \log \frac{1}{1-z} \right)^{\beta}, \tag{A.86}$$

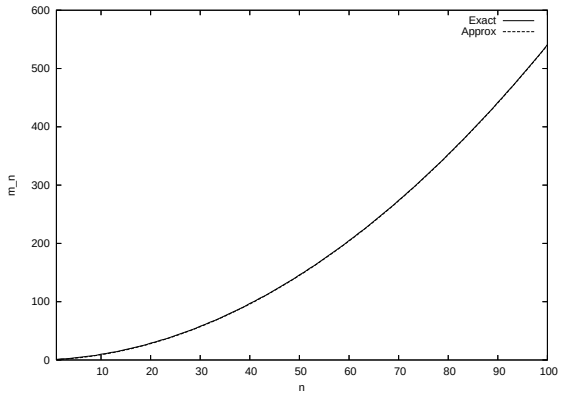


Figure A.4: The comparison of the exact and approximative solutions for the money changing problem for $n = 1, \ldots, 100$.

for real numbers α, β. For the purposes of the other parts of this thesis, however, the following special version is more appropriate: Let $\alpha \neq 0, -1, -2, \ldots$. Then the coefficient of z^n in $(1-z)^{-\alpha}$ is given by

$$[z^n](1-z)^{-\alpha} \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)} \left(1 + \sum_{k=1}^{\infty} \frac{e_k(\alpha)}{n^k}\right), \tag{A.87}$$

where $e_k(\alpha)$ is a polynomial in α of degree $2k$. The first few polynomials are given by

$$e_1(\alpha) = \frac{\alpha(\alpha-1)}{2} \tag{A.88}$$

$$e_2(\alpha) = \frac{\alpha(\alpha-1)(\alpha-2)(3\alpha-1)}{24} \tag{A.89}$$

$$e_3(\alpha) = \frac{\alpha^2(\alpha-1)^2(\alpha-2)(\alpha-3)}{48}. \tag{A.90}$$

The exact definition of these polynomials is complicated but can be found in [66].

To illustrate the use of (A.87), we show how to calculate the asymptotic form for the coefficients of $(1-az)^{-1/2}$, where a is a constant. Firstly, we notice a simple fact that

$$[z^n](1-az)^{-\alpha} = a^n[z^n](1-z)^{-\alpha}. \tag{A.91}$$

Function	Coefficients
$(1-z)^{3/2}$	$\frac{1}{\sqrt{\pi n^5}}\left(\frac{3}{4}+\frac{45}{32n}+\frac{1155}{512n^2}+\mathcal{O}\left(\frac{1}{n^3}\right)\right)$
$(1-z)$	0
$(1-z)^{1/2}$	$-\frac{1}{\sqrt{\pi n^3}}\left(\frac{1}{2}+\frac{3}{16n}+\frac{25}{256n^2}+\mathcal{O}\left(\frac{1}{n^3}\right)\right)$
1	0
$(1-z)^{-1/2}$	$\frac{1}{\sqrt{\pi n}}\left(1-\frac{1}{8n}+\frac{1}{128n^2}+\mathcal{O}\left(\frac{1}{n^3}\right)\right)$
$(1-z)^{-1}$	1
$(1-z)^{-3/2}$	$\sqrt{\frac{n}{\pi}}\left(2+\frac{3}{4n}-\frac{7}{64n^2}+\mathcal{O}\left(\frac{1}{n^3}\right)\right)$
$(1-z)^{-2}$	$n+1$
$(1-z)^{-3}$	$\frac{1}{2}n^2+\frac{3}{2}n+1$
$(1-z)^{-4}$	$\frac{1}{6}n^3+n^2+\frac{11}{6}n+1$

Table A.2: Some commonly encountered functions and the asymptotic form of their coefficients.

The value of α in our example is $1/2$, so

$$[z^n](1-az)^{-1/2} \sim a^n \cdot \frac{n^{-1/2}}{\Gamma(1/2)}\left[1+\frac{(1/2)(-1/2)}{2n}\right. \tag{A.92}$$
$$\left.+\frac{(1/2)(-1/2)(-3/2)(1/2)}{24n^2}+\mathcal{O}\left(\frac{1}{n^3}\right)\right]$$
$$= a^n \cdot \frac{1}{\sqrt{\pi n}}\left(1-\frac{1}{8n}+\frac{1}{128n^2}+\mathcal{O}\left(\frac{1}{n^3}\right)\right). \tag{A.93}$$

Further examples are listed in Table A.2.

Another very important result of singularity analysis is the following *transfer theorem*: If a generating function $A(z)$ satisfies

$$A(z) = \mathcal{O}\left((1-z)^{-\alpha}\right), \tag{A.94}$$

then

$$[z^n]A(z) = \mathcal{O}\left(n^{\alpha-1}\right). \tag{A.95}$$

The same holds for the $o(\cdot)$-functions. Comparing the transfer theorem (A.95) to Equation (A.87), we can see that it is actually very intuitive.

We finalize this section by summarizing the method of singularity analysis into the following recipe:

1. Find the generating function $A(z)$ for the sequence we are interested in.
2. Find the dominant singularity of $A(z)$.
3. Expand $A(z)$ into series around the dominant singularity.
4. Apply Theorems (A.87) and (A.95) to get the asymptotic form for the coefficients.

A highly non-trivial example of using this recipe is presented in Appendix B.

Chapter B

The Szpankowski Approximation

In this appendix we will first derive a generating function for the sequence of multinomial regret terms. This function is used twice in the other parts of this thesis: The elegant recursion formula for exact NML computation in Section 3.1.1 and the accurate Szpankowski approximation in Section 3.1.2 are based on this generating function. Secondly, we give full derivation of the Szpankowski approximation.

B.1 The Regret Generating Function

Let us start with the sequence $\langle n^n/n! \rangle$. As in [66], we denote the function generating this sequence by $B(z)$. Unfortunately, there is no closed-form formula for $B(z)$. As we will see later, this function is nevertheless suitable for our purposes. The connection between $B(z)$ and the multinomial regret

terms can be seen by squaring $B(z)$,

$$B^2(z) = \left(\sum_{r\geq 0} \frac{r^r}{r!} z^r\right) \cdot \left(\sum_{s\geq 0} \frac{s^s}{s!} z^s\right) \tag{B.1}$$

$$= \sum_{r,s\geq 0} \frac{r^r s^s}{r!s!} z^{r+s} \tag{B.2}$$

$$= \sum_{n\geq 0} \left(\sum_{r+s=n} \frac{n^n}{n!} \frac{n!}{r!s!} \frac{r^r s^s}{n^{r+s}}\right) z^n \tag{B.3}$$

$$= \sum_{n\geq 0} \frac{n^n}{n!} \left(\sum_{r+s=n} \frac{n!}{r!s!} \left(\frac{r}{n}\right)^r \left(\frac{s}{n}\right)^s\right) z^n \tag{B.4}$$

$$= \sum_{n\geq 0} \frac{n^n}{n!} \mathcal{C}(\mathcal{M}(2), n) z^n, \tag{B.5}$$

where $\mathcal{M}(2)$ is the multinomial model class with two values. Thus, $B^2(z)$ generates the sequence $\langle \frac{n^n}{n!} \mathcal{C}(\mathcal{M}(2), n)\rangle$. This easily generalizes to

$$B^K(z) = \sum_{n\geq 0} \frac{n^n}{n!} \left[\sum_{h_1+\cdots+h_K=n} \frac{n!}{h_1!\cdots h_K!} \left(\frac{h_1}{n}\right)^{h_1} \cdots \left(\frac{h_k}{n}\right)^{h_k}\right] z^n \tag{B.6}$$

$$= \sum_{n\geq 0} \frac{n^n}{n!} \mathcal{C}(\mathcal{M}(K), n) z^n, \tag{B.7}$$

generating the sequence $\langle \frac{n^n}{n!} \mathcal{C}(\mathcal{M}(K), n)\rangle$. Note that to be precise, the function $B^K(z)$ is the *tree-like generating function* [66] of the sequence $\langle \mathcal{C}(\mathcal{M}(K), n)\rangle$. For simplicity, however, we just call it the regret generating function.

To make the Equation (B.7) useful, we will derive a relation of $B^K(z)$ and the so-called *Cayley's tree function* $T(z)$ [30, 9], which generates the sequence $\langle n^{n-1}/n!\rangle$, i.e.,

$$T(z) = \sum_{n\geq 1} \frac{n^{n-1}}{n!} z^n, \tag{B.8}$$

as shown in [66]. This sequence counts the *rooted labeled trees*, hence the name of the function. The tree function is defined by the functional equation

$$T(z) = z e^{T(z)}. \tag{B.9}$$

Differentiating and multiplying (B.8) by z, we get

$$zT'(z) = z \cdot \sum_{n \geq 1} \frac{n \cdot n^{n-1}}{n!} z^{n-1} \tag{B.10}$$

$$= \sum_{n \geq 1} \frac{n^n}{n!} z^n \tag{B.11}$$

$$= \sum_{n \geq 0} \frac{n^n}{n!} z^n - 1, \tag{B.12}$$

from which we get

$$B(z) = zT'(z) + 1. \tag{B.13}$$

On the other hand, differentiating the functional equation (B.9) gives

$$T'(z) = e^{T(z)} + ze^{T(z)} \cdot T'(z) \tag{B.14}$$

$$T'(z)(1 - ze^{T(z)}) = e^{T(z)} \tag{B.15}$$

$$zT'(z)(1 - T(z)) = T(z) \tag{B.16}$$

$$zT'(z) = \frac{T(z)}{1 - T(z)}. \tag{B.17}$$

Combining the Equations (B.13) and (B.17), we get

$$B(z) = \frac{T(z)}{1 - T(z)} + 1 = \frac{1}{1 - T(z)}, \tag{B.18}$$

and thus

$$B^K(z) = \frac{1}{(1 - T(z))^K}. \tag{B.19}$$

This final form can now applied in NML computation by using the properties of the tree function $T(z)$.

B.2 The Derivation

The proof of the Szpankowski approximation (3.8) was only outlined in [66]. We will now present a full derivation. Our starting point is the regret generating function already discussed in Appendix B.1,

$$B^K(z) = \frac{1}{(1 - T(z))^K} = \sum_{n \geq 0} \frac{n^n}{n!} \mathcal{C}(\mathcal{M}(K), n) z^n. \tag{B.20}$$

To make the presentation easier to follow, the derivation is split into the following steps:

1. Find the dominant singularity of the regret generating function $B^K(z)$.
2. Expand the inverse of the tree function $T(z)$ into series around the dominant singularity point.
3. Invert this series to get the expansion of the tree function.
4. Find the series for $B(z) = 1/(1 - T(z))$.
5. Find the series for $B^K(z)$.
6. Apply the singularity analysis theorem (A.87) term by term.
7. Multiply by $n!/n^n$ to extract the asymptotic form of the regret terms.
8. Take the logarithm to prove (3.8).

Step 1: To get the asymptotic form for the coefficients of (B.20), we need to expand the function $B^K(z)$ around its dominant singularity, i.e., the one nearest to the origo. It is well-known (see, e.g., [9]) that the dominant singularity of $T(z)$ occurs at $z = 1/e$. This point is also the dominant singularity of (B.20), since the zero of the denominator (pole) is also at $z = 1/e$. This can be seen by solving z from the functional equation (B.9),

$$z = F(T) = Te^{-T}, \tag{B.21}$$

and then plugging $T = 1$ into it.

Step 2: Deriving the series expansion for (B.20) is a very non-trivial task, since there is no explicit formula for $B(z)$ or $T(z)$. It turns out that the inverse function $F(T)$ is a good starting point, since it is an entire function (analytic everywhere). To get the expansion of $T(z)$ around $z = 1/e$, we can first expand $F(T)$ around $T = 1$, and then use the series inversion method described in Appendix A.2.5. Since $F(T)$ is entire, its expansion is a simple Taylor series, which can be found by calculating the derivatives of $F(T)$ at $T = 1$. We have

$$\begin{aligned}
F'(T) &= e^{-T} + T \cdot (-e^{-T}) = e^{-T}(1 - T) && \text{(B.22)}\\
F''(T) &= -e^{-T}(1 - T) - e^{-T} = -e^{-T}(2 - T) && \text{(B.23)}\\
F'''(T) &= e^{-T}(2 - T) + e^{-T} = e^{-T}(3 - T) && \text{(B.24)}\\
F''''(T) &= -e^{-T}(3 - T) - e^{-T} = -e^{-T}(4 - T), && \text{(B.25)}
\end{aligned}$$

which leads to

$$F(T) = F(1) + F'(1)(T-1) + \frac{F''(1)}{2!}(T-1)^2 + \frac{F'''(1)}{3!}(T-1)^3 + \frac{F''''(1)}{4!}(T-1)^4 + \cdots \quad \text{(B.26)}$$

$$= 1/e - \frac{1/e}{2}(T-1)^2 + \frac{1/e}{3}(T-1)^3 - \frac{1/e}{8}(T-1)^4 + \cdots \quad \text{(B.27)}$$

$$= 1/e - \frac{1/e}{2}(1-T)^2 - \frac{1/e}{3}(1-T)^3 - \frac{1/e}{8}(1-T)^4 + \cdots . \quad \text{(B.28)}$$

Step 3: Looking at Equation (B.22), we can see that the first derivative vanishes at $T = 1$. As suggested in Appendix A.2.5, this unfortunately means that inverting the series (B.28) is not straightforward. Intuitively, this complication can be understood via Figure B.1, where the function $F(T)$ is plotted near the point $T = 1$ (in real number space). Clearly, $F(T)$ is non-monotonic in every neighborhood of $T = 1$, and the

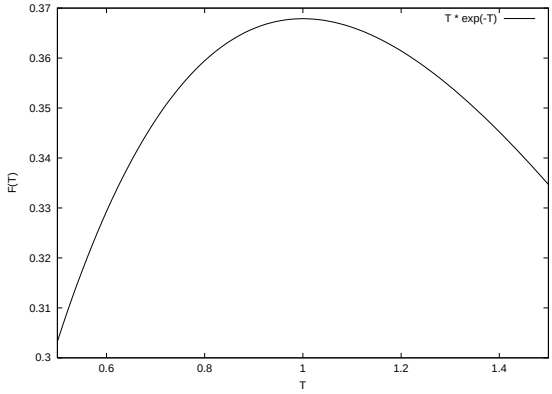


Figure B.1: Plot of $F(T) = Te^{-T}$ around $T = 1$.

inverse function thus multiple-valued. It follows that the expansion of $T(z)$ around point $z = 1/e$ must also contain multiple-valued terms. As we will soon see, this is indeed the case: the inverted series will be a *Puiseux series* with fractional power terms. To read more about Puiseux series, see Appendix A.1.7.

To find the inverse of (B.28), we can use a theorem from [14], which classifies series expansions into four types of systematic patterns based on

the first few terms of the series. With the terminology of [14], our series falls into category "Type II" with the order parameter β set to 2 (see also Appendix A.1.7). For this category, the series inversion is performed by starting with variable transformations

$$v = 1 - T \tag{B.29}$$

$$w = (1/e - F(T))^{1/\beta} = (1/e - z)^{1/2}, \tag{B.30}$$

and then examining the function

$$w = A(v) = (1/e - F(T))^{1/2} \tag{B.31}$$

$$= (f_2 v^2 + f_3 v^3 + f_4 v^4 + \cdots)^{1/2}, \tag{B.32}$$

where, from (B.28),

$$f_2 = \frac{1/e}{2}, \quad f_3 = \frac{1/e}{3}, \quad f_4 = \frac{1/e}{8}. \tag{B.33}$$

Next we need to find the series expansion for function $A(v)$, i.e., coefficients s_n such that

$$\left(f_2 v^2 + f_3 v^3 + f_4 v^4 + \cdots\right)^{1/2} = s_1 v + s_2 v^2 + s_3 v^3 + \cdots . \tag{B.34}$$

It is easy to prove (see also Figure B.1) that $1/e - F(T) \geq 0$ for all $T \in \mathbb{R}$, from which it follows that we can square both sides of (B.34)

$$f_2 v^2 + f_3 v^3 + f_4 v^4 + \cdots = \left(s_1 v + s_2 v^2 + s_3 v^3 + \cdots\right)^2 \tag{B.35}$$

$$= s_1^2 v^2 + 2 s_1 s_2 v^3 + (2 s_1 s_3 + s_2^2) v^4 + \cdots , \tag{B.36}$$

and by coefficient comparison

$$s_1^2 = f_2, \quad s_1 = \sqrt{f_2} \tag{B.37}$$

$$2 s_1 s_2 = f_3, \quad s_2 = \frac{f_3}{2 s_1} = \frac{f_3}{2\sqrt{f_2}} \tag{B.38}$$

$$2 s_1 s_3 + s_2^2 = f_4, \quad s_3 = \frac{f_4 - s_2^2}{2 s_1} = \frac{4 f_2 f_4 - f_3^2}{8 f_2^{3/2}}. \tag{B.39}$$

The function $A(v)$ can now be written as

$$A(v) = \sqrt{f_2} v + \frac{f_3}{2\sqrt{f_2}} v^2 + \frac{4 f_2 f_4 - f_3^2}{8 f_2^{3/2}} v^3 + \cdots , \tag{B.40}$$

from which we can finally see the idea behind the transformations (B.29) and (B.30). That is, series (B.40) is an ordinary power series with zero constant coefficient therefore having a well-defined inverse, say,

$$v = D(w) = d_1 w + d_2 w^2 + d_3 w^3 + \cdots , \tag{B.41}$$

where the coefficients d_n are given by (see Appendix A.2.5)

$$d_1 = \frac{1}{s_1} = \frac{1}{\sqrt{f_2}} = \sqrt{2e} \tag{B.42}$$

$$d_2 = -\frac{s_2}{s_1^3} = -\frac{f_3}{2f_2^2} = -\frac{2}{3}e \tag{B.43}$$

$$d_3 = \frac{2s_2^2}{s_1^5} - \frac{s_3}{s_1^4} = \frac{5f_3^2 - 4f_2 f_4}{8f_2^{7/2}} = \frac{11\sqrt{2}}{36}e^{3/2}. \tag{B.44}$$

Transforming back to original variables gives the series expansion for the tree function

$$T(z) = 1 - D(w) \tag{B.45}$$

$$= 1 - \sqrt{2e}(1/e - z)^{1/2} + \frac{2}{3}e(1/e - z) - \frac{11\sqrt{2}}{36}e^{3/2}(1/e - z)^{3/2} + \cdots , \tag{B.46}$$

which can be further written as

$$T(z) = 1 - \sqrt{2}(1 - ez)^{1/2} + \frac{2}{3}(1 - ez) - \frac{11\sqrt{2}}{36}(1 - ez)^{3/2} + \cdots . \tag{B.47}$$

This final form makes is more convenient to apply singularity analysis in Step 6.

Step 4: After deriving the expansion for $T(z)$, the next task is to find series for

$$B(z) = \frac{1}{1 - T(z)}, \tag{B.48}$$

i.e., the reciprocal series of

$$1 - T(z) = \sqrt{2}(1 - ez)^{1/2} - \frac{2}{3}(1 - ez) + \frac{11\sqrt{2}}{36}(1 - ez)^{3/2} + \cdots . \tag{B.49}$$

It is clear that the reciprocal is of the form

$$B(z) = a(1 - ez)^{-1/2} + b + c(1 - ez)^{1/2} + \cdots , \tag{B.50}$$

for some numbers $(a, b, c, \ldots)$. By the definition of the reciprocal series, we must then have

$$B(z)(1-T(z)) = \left[a(1-ez)^{-1/2} + b + c(1-ez)^{1/2} + \cdots\right] \cdot \left[\sqrt{2}(1-ez)^{1/2} - \frac{2}{3}(1-ez) + \frac{11\sqrt{2}}{36}(1-ez)^{3/2} + \cdots\right] \equiv 1, \quad \text{(B.51)}$$

i.e., the trivial sequence $(1, 0, 0, \ldots)$. The coefficients $(a, b, c, \ldots)$ can be calculated by comparing coefficients

$$\sqrt{2}a = 1, \quad a = \frac{1}{\sqrt{2}} \quad \text{(B.52)}$$

$$-\frac{2}{3}a + \sqrt{2}b = 0, \quad b = \frac{2}{3\sqrt{2}}a = \frac{1}{3} \quad \text{(B.53)}$$

$$\frac{11\sqrt{2}}{36}a - \frac{2}{3}b + \sqrt{2}c = 0, \quad c = -\frac{11}{36}a + \frac{2}{3\sqrt{2}}b = -\frac{\sqrt{2}}{24}, \quad \text{(B.54)}$$

and thus we get the series expansion

$$B(z) = \frac{1}{\sqrt{2}}(1-ez)^{-1/2} + \frac{1}{3} - \frac{\sqrt{2}}{24}(1-ez)^{1/2} + \cdots. \quad \text{(B.55)}$$

Step 5: The final step for deriving the series expansion for the regret generating function (B.20) is to expand

$$B^K(z) = \frac{1}{(1-T(z))^K} = \left(\frac{1}{\sqrt{2}}(1-ez)^{-1/2} + \frac{1}{3} - \frac{\sqrt{2}}{24}(1-ez)^{1/2} + \cdots\right)^K. \quad \text{(B.56)}$$

The first term of this series, i.e., the one with the smallest exponent, is obtained by raising the first term of (B.56) into Kth power

$$\left(\frac{1}{\sqrt{2}}(1-ez)^{-1/2}\right)^K = \left(\frac{1}{\sqrt{2}}\right)^K (1-ez)^{-K/2} = \frac{1}{2^{K/2}}(1-ez)^{-K/2}. \quad \text{(B.57)}$$

To get the next term we raise the first term of (B.56) into $(K-1)$th power and then multiply by the second term. There are K different ways to choose the second term, which gives

$$K \cdot \left(\frac{1}{\sqrt{2}}\right)^{K-1} \cdot \frac{1}{3} \cdot (1-ez)^{-\frac{K}{2}+\frac{1}{2}} = \frac{K}{3 \cdot 2^{\frac{K}{2}-\frac{1}{2}}}(1-ez)^{-\frac{K}{2}+\frac{1}{2}}. \quad \text{(B.58)}$$

For the third term, we need to consider two cases:

1. Raise the first term of (B.56) into $(K-1)$th power and then multiply by the third term. The third term can be chosen in K different ways.

2. Raise the first term of (B.56) into $(K-2)$th power and then multiply by the square of the second term. We have $\binom{K}{2} = K(K-1)/2$ ways to do that.

Thus, the third term of $B^K(z)$ is

$$\left[K \cdot \left(\frac{1}{\sqrt{2}}\right)^{K-1} \cdot \frac{-\sqrt{2}}{24} + \frac{K(K-1)}{2} \cdot \left(\frac{1}{\sqrt{2}}\right)^{K-2} \cdot \left(\frac{1}{3}\right)^2\right] \cdot (1-ez)^{-\frac{K}{2}+1} = \frac{4K(K-1)-3K}{36 \cdot 2^{K/2}}(1-ez)^{-\frac{K}{2}+1}. \quad \text{(B.59)}$$

As we will soon see, it is not necessary to calculate more terms. The series expansion for the regret generating function is now

$$B^K(z) = \frac{1}{2^{K/2}}(1-ez)^{-K/2} + \frac{K}{3 \cdot 2^{\frac{K}{2}-\frac{1}{2}}}(1-ez)^{-\frac{K}{2}+\frac{1}{2}} + \frac{4K(K-1)-3K}{36 \cdot 2^{K/2}}(1-ez)^{-\frac{K}{2}+1} + \cdots . \quad \text{(B.60)}$$

Step 6: We are now ready to apply the singularity analysis theorem (A.87) to series (B.60). Proceeding term by term basis,

$$[z^n]\left(\frac{1}{2^{K/2}}(1-ez)^{-K/2}\right) \sim \quad \text{(B.61)}$$
$$e^n \cdot \frac{n^{\frac{K}{2}-1}}{2^{K/2} \cdot \Gamma(K/2)}\left(1 + \frac{K(K-1)}{2n} + \mathcal{O}\left(1/n^2\right)\right)$$
$$[z^n]\left(\frac{K}{3 \cdot 2^{\frac{K}{2}-\frac{1}{2}}}(1-ez)^{-\frac{K}{2}+\frac{1}{2}}\right) \sim \quad \text{(B.62)}$$
$$e^n \cdot \frac{K \cdot n^{\frac{K}{2}-\frac{3}{2}}}{3 \cdot 2^{\frac{K}{2}-\frac{1}{2}} \cdot \Gamma(\frac{K}{2}-\frac{1}{2})}\left(1 + \frac{K(K-1)}{2n} + \mathcal{O}\left(1/n^2\right)\right)$$
$$[z^n]\left(\frac{4K(K-1)-3K}{36 \cdot 2^{K/2}}(1-ez)^{-\frac{K}{2}+1}\right) \sim \quad \text{(B.63)}$$
$$e^n \cdot \frac{(4K(K-1)-3K) \cdot n^{\frac{K}{2}-2}}{36 \cdot 2^{K/2} \cdot \Gamma(\frac{K}{2}-1)}\left(1 + \frac{K(K-1)}{2n} + \mathcal{O}\left(1/n^2\right)\right).$$

After some tedious algebra we get the asymptotic form for the nth coefficient of the regret generating function:

$$[z^n]B^K(z) \sim e^n \cdot \left[\frac{1}{2^{K/2} \cdot \Gamma(K/2)} \cdot n^{\frac{K}{2}-1} + \frac{K}{2^{\frac{K}{2}-\frac{1}{2}} \cdot 3\Gamma(\frac{K}{2}-\frac{1}{2})} \cdot n^{\frac{K}{2}-\frac{3}{2}} + \frac{K(K-2)(2K+1)}{2^{K/2} \cdot 36\Gamma(K/2)} \cdot n^{\frac{K}{2}-2} + \mathcal{O}\left(n^{\frac{K}{2}-\frac{5}{2}}\right) \right]. \quad \text{(B.64)}$$

Step 7: To extract the asymptotic form of the terms $\mathcal{C}(\mathcal{M}(K), n)$, we need to multiply Equation (B.64) by $n!/n^n$. By the celebrated Stirling's formula,

$$\frac{n!}{n^n} = \sqrt{2\pi n} \cdot e^{-n} \left(1 + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^2}\right)\right), \quad \text{(B.65)}$$

which nicely cancels the e^n term in (B.64). Multiplying (B.64) by (B.65) gives after simplifications

$$\begin{aligned} \mathcal{C}(\mathcal{M}(K), n) &\sim \left(\frac{n}{2}\right)^{\frac{K-1}{2}} \cdot \frac{\sqrt{\pi}}{\Gamma(K/2)} \left[1 + \frac{\sqrt{2}K \cdot \Gamma(K/2)}{3\Gamma(\frac{K}{2}-\frac{1}{2})} \cdot \frac{1}{\sqrt{n}} \right. \quad \text{(B.66)} \\ &\quad \left. + \frac{K(K-2)(2K+1)}{36} \cdot \frac{1}{n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right] \\ &\quad \cdot \left[1 + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^2}\right)\right] \\ &= \left(\frac{n}{2}\right)^{\frac{K-1}{2}} \cdot \frac{\sqrt{\pi}}{\Gamma(K/2)} \left[1 + \frac{\sqrt{2}K \cdot \Gamma(K/2)}{3\Gamma(\frac{K}{2}-\frac{1}{2})} \cdot \frac{1}{\sqrt{n}} \right. \quad \text{(B.67)} \\ &\quad \left. + \frac{3 + K(K-2)(2K+1)}{36} \cdot \frac{1}{n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right]. \end{aligned}$$

Step 8: The final step is to take the logarithm of (B.67). Consider the standard Taylor series of the (natural) logarithm function

$$\log(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} + \cdots. \quad \text{(B.68)}$$

Plugging

$$z = \frac{a}{\sqrt{n}} + \frac{b}{n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right) \quad \text{(B.69)}$$

into series (B.68) gives

$$\log\left[1+\frac{a}{\sqrt{n}}+\frac{b}{n}+\mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right]=\frac{a}{\sqrt{n}}+\frac{b}{n}-\frac{1}{2}\left[\frac{a}{\sqrt{n}}+\frac{b}{n}+\mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right]^2 \tag{B.70}$$

$$=\frac{a}{\sqrt{n}}+(b-\frac{1}{2}a^2)\cdot\frac{1}{n}+\mathcal{O}\left(\frac{1}{n^{3/2}}\right), \tag{B.71}$$

for numbers a, b. By applying (B.71) to (B.67) we get the asymptotic formula for the multinomial regret terms:

$$\log\mathcal{C}(\mathcal{M}(K),n)=\frac{K-1}{2}\log\frac{n}{2}+\log\frac{\sqrt{\pi}}{\Gamma(K/2)}+\frac{\sqrt{2}K\cdot\Gamma(K/2)}{3\Gamma(\frac{K}{2}-\frac{1}{2})}\cdot\frac{1}{\sqrt{n}} \tag{B.72}$$

$$+\left(\frac{3+K(K-2)(2K+1)}{36}-\frac{\Gamma^2(K/2)\cdot K^2}{9\Gamma^2\left(\frac{K}{2}-\frac{1}{2}\right)}\right)\cdot\frac{1}{n} \tag{B.73}$$

$$+\mathcal{O}\left(\frac{1}{n^{3/2}}\right).$$

The proof of (3.8) follows trivially.

An important thing to notice is that in all the steps of the derivation we could have calculated an arbitrary number of terms for the series expansions. It follows that the derivation does not limit the accuracy of the final result. However, as shown in Section 3.1.3, $\mathcal{O}\left(1/n^{3/2}\right)$ is accurate enough for practical purposes.

References

[1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, 1970.

[2] A. Asuncion and D. Newman. UCI machine learning repository, 2007. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[3] V. Balakrishnan. *Schaum's Outline of Theory and Problems of Combinatorics*. McGraw-Hill, 1995.

[4] V. Balasubramanian. MDL, Bayesian inference, and the geometry of the space of probability distributions. In P. Grünwald, I. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, pages 81–98. The MIT Press, 2006.

[5] A. Barron, J. Rissanen, and B. Yu. The minimum description principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, October 1998.

[6] L. Birge and Y. Rozenholc. How many bins should be put in a regular histogram. Prepublication no 721, Laboratoire de Probabilites et Modeles Aleatoires, CNRS-UMR 7599, Universite Paris VI & VII, April 2002.

[7] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman. Autoclass: A Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 54–64, Ann Arbor, June 1988.

[8] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[9] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5:329–359, 1996.

[10] N. De Bruijn. *Asymptotic Methods in Analysis*. Dover Publications, Inc., New York, 1981.

[11] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[12] E. Elovaara and P. Myllymäki. MDL-based attribute models in naive Bayes classification. In *Workshop on Information Theoretic Methods in Science and Engineering (WITMSE)*, Tampere, Finland, 2009.

[13] B. Everitt and D. Hand. *Finite Mixture Distributions*. Chapman and Hall, London, 1981.

[14] B. Fabinojas. Laplace's method on a computer algebra system with an application to the real valued modified Bessel functions. *Journal of Computational and Applied Mathematics*, 146:323–342, 2002.

[15] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, 3rd edition, 1968.

[16] P. Flajolet and A. Odlyzko. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, 3(2):216–240, 1990.

[17] P. Flajolet and R. Sedgewick. The average case analysis of algorithms : Complex asymptotics and generating functions. Technical Report RR-2026, INRIA, 1993.

[18] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[19] R. Graham, D. Knuth, and O. Patashnik. *Concrete Mathematics (second edition)*. Addison-Wesley, 1994.

[20] D. Greene and D. Knuth. *Mathematics for the Analysis of Algorithms*. Birkhäuser Boston, 1982.

[21] P. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[22] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Minimum encoding approaches for predictive modeling. In G. Cooper and S. Moral, editors, *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 183–192,

Madison, WI, July 1998. Morgan Kaufmann Publishers, San Francisco, CA.

[23] P. Hall and E. Hannan. On stochastic complexity and nonparametric density estimation. *Biometrika*, 75(4):705–714, 1988.

[24] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, One Microsoft Way, Redmond, WA 98052, 1996.

[25] P. Henrici. Automatic computations with power series. *Journal of the ACM*, 3(1):11–15, January 1956.

[26] P. Henrici. *Applied and Computational Complex Analysis, Vols. 1–3.* John Wiley & Sons, New York, 1977.

[27] D. Knuth. *The Art of Computer Programming, vol. 1 / Fundamental Algorithms (third edition).* Addison-Wesley, 1997.

[28] D. Knuth. *The Art of Computer Programming, vol. 2 / Seminumerical Algorithms (third edition).* Addison-Wesley, 1998.

[29] D. Knuth. *The Art of Computer Programming, vol. 3 / Sorting and Searching (second edition).* Addison-Wesley, 1998.

[30] D. Knuth and B. Pittel. A recurrence related to trees. *Proceedings of the American Mathematical Society*, 105(2):335–349, 1989.

[31] M. Koivisto. *Sum-Product Algorithms for the Analysis of Genetic Risks.* PhD thesis, Report A-2004-1, Department of Computer Science, University of Helsinki, 2004.

[32] P. Kontkanen, W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri. Efficient computation of stochastic complexity. In C. Bishop and B. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics.* Society for Artificial Intelligence and Statistics, 2003.

[33] P. Kontkanen, J. Lahtinen, P. Myllymäki, T. Silander, and H. Tirri. Supervised model-based visualization of high-dimensional data. *Intelligent Data Analysis*, 4:213–227, 2000.

[34] P. Kontkanen and P. Myllymäki. A fast normalized maximum likelihood algorithm for multinomial data. In L. P. Kaelbling and A. Saffiotti, editors, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, 2005.

[35] P. Kontkanen and P. Myllymäki. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103(6):227–233, 2007.

[36] P. Kontkanen and P. Myllymäki. MDL histogram density estimation. In M. Meila and S. Shen, editors, *Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics*, March 2007.

[37] P. Kontkanen and P. Myllymäki. An empirical comparison of NML clustering algorithms. In M. Dehmer, M. Drmota, and F. Emmert-Streib, editors, *Proceedings of the International Conference on Information Theory and Statistical Learning (ITSL-08)*. CSREA Press, 2008.

[38] P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri. An MDL framework for data clustering. In P. Grünwald, I. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*. The MIT Press, 2005.

[39] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On Bayesian case matching. In B. Smyth and P. Cunningham, editors, *Advances in Case-Based Reasoning, Proceedings of the 4th European Workshop (EWCBR-98)*, volume 1488 of *Lecture Notes in Artificial Intelligence*, pages 13–24. Springer-Verlag, 1998.

[40] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. On predictive distributions and Bayesian networks. *Statistics and Computing*, 10:39–54, 2000.

[41] P. Kontkanen, P. Myllymäki, and H. Tirri. Constructing Bayesian finite mixture models by the EM algorithm. Technical Report NC-TR-97-003, ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT), 1996.

[42] P. Kontkanen, H. Wettig, and P. Myllymäki. NML computation algorithms for tree-structured multinomial Bayesian networks. *EURASIP Journal on Bioinformatics and Systems Biology*, Article ID 90947, 2007.

[43] G. Korodi and I. Tabus. An efficient normalized maximum likelihood algorithm for DNA sequence compression. *ACM Trans. Inf. Syst.*, 23(1):3–34, 2005.

[44] G. McLachlan, editor. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.

[45] M. Meila and D. Heckerman. An experimental comparison of several clustering and initialization methods. In G. F. Cooper and S. Moral, editors, *UAI'98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 386–395, 1998.

[46] T. Mononen and P. Myllymäki. Fast NML computation for Naive Bayes models. In V. Corruble, M. Takeda, and E. Suzuki, editors, *Proceedings of the Tenth International Conference on Discovery Science*, October 2007.

[47] T. Mononen and P. Myllymäki. Computing the multinomial stochastic complexity in sub-linear time. In *Proceedings of European Workshop on Probabilistic Graphical Models (PGM'08)*, pages 209–216, 2008.

[48] T. Mononen and P. Myllymäki. Computing the NML for Bayesian forests via matrices and generating polynomials. In *IEEE Information Theory Workshop*, Porto, Portugal, May 2008.

[49] T. Mononen and P. Myllymäki. On recurrence formulas for computing the stochastic complexity. In *Proceedings of the International Symposium on Information Theory and its Applications*, pages 281–286, Auckland, New Zealand, 2008. IEEE.

[50] T. Mononen and P. Myllymäki. On the multinomial stochastic complexity and its connection to the birthday problem. In *International Conference on Information Theory and Statistical Learning*, Las Vegas, NV, July 2008.

[51] T. Needham. *Visual Complex Analysis*. Oxford University Press, 1997.

[52] A. Odlyzko. Asymptotic enumeration methods. In R. L. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics*, volume 2, pages 1063–1229. North-Holland, Amsterdam, 1995.

[53] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:445–471, 1978.

[54] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society*, 49(3):223–239 and 252–265, 1987.

[55] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, New Jersey, 1989.

[56] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January 1996.

[57] J. Rissanen. Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47(5):1712–1717, July 2001.

[58] J. Rissanen. *Information and Complexity in Statistical Modeling*. Springer, 2007.

[59] J. Rissanen, T. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2):315–323, March 1992.

[60] T. Roos, P. Myllymäki, and H. Tirri. On the behavior of MDL denoising. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, pages 309–316, 2005.

[61] T. Roos, T. Silander, P. Kontkanen, and P. Myllymäki. Bayesian network structure learning using factorized NML universal models. In *Information Theory and Applications Workshop*, San Diego, CA, January 2008.

[62] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[63] Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23:3–17, 1987.

[64] P. Smyth. Probabilistic model-based clustering of multivariate and sequential data. In D. Heckerman and J. Whittaker, editors, *Proceedings of the Seventh International Conference on Artificial Intelligence and Statistics*, pages 299–304. Morgan Kaufmann Publishers, 1999.

[65] M. Spiegel. *Schaum's Outline of Theory and Problems of Complex Variables*. McGraw-Hill, 1981.

[66] W. Szpankowski. *Average case analysis of algorithms on sequences*. John Wiley & Sons, 2001.

[67] I. Tabus, J. Rissanen, and J. Astola. Classification and feature gene selection using the normalized maximum likelihood model for discrete regression. *Signal Processing, Special issue on Genomic Signal Processing*, 83(4):713–727, 2003.

[68] H. Tirri. *Plausible Prediction by Bayesian Inference*. PhD thesis, Report A-1997-1, Department of Computer Science, University of Helsinki, June 1997.

[69] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York, 1985.

[70] H. Wettig, P. Kontkanen, and P. Myllymäki. Calculating the normalized maximum likelihood distribution for Bayesian forests. *IADIS International Journal on Computer Science and Information Systems*, 2, October 2007.

[71] H. Wilf. *generatingfunctionology (second edition)*. Academic Press, 1994.

[72] Q. Xie and A. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, March 2000.

[73] B. Yu and T. Speed. Data compression and histograms. *Probab. Theory Relat. Fields*, 92:195–229, 1992.

[74] D. Zill and P. Shanahan. *A First Course in Complex Analysis with Applications*. Jones and Bartlett Publishers, Inc., 2003.