The Minimum Description Length Principle and Reasoning under Uncertainty

Peter Grünwald

Warning: This is the web-version of the thesis. The original version of this thesis was printed as a book by a professional printer. It was set in the *Lucida* font. Because of technical difficulties, the present on-line version had to be set in the standard Latex (Computer Modern) Font. Since the lay-out of the thesis had been specifically designed for the beautiful Lucida font, the present version's lay-out is non-optimal. In order to obtain the book version, please contact the author at `grunwald@cs.stanford.edu`.

ILLC Dissertation Series 1998-03

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
phone: +31-20-5256090
fax: +31-20-5255101
e-mail: illc@wins.uva.nl

The Minimum Description Length Principle and Reasoning under Uncertainty

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam,
op gezag van de Rector Magnificus
prof.dr. J.J.M. Franse
ten overstaan van een door het college voor promoties
ingestelde commissie in het openbaar te verdedigen in de
Aula der Universiteit
op donderdag 8 oktober 1998 te 11.00 uur

door

Peter Daniel Grünwald

geboren te Geldrop.

Promotor: Prof.dr.ir. P.M.B. Vitányi

Faculteit Wiskunde, Informatica, Natuurkunde en Sterrenkunde
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam

The research reported in this thesis was done at the CWI (Centrum voor Wiskunde en Informatica), the National Centre for Mathematics and Computer Science. The investigations were supported by the Computer Science Research Foundation (SION) which is subsidized by the Netherlands Organization for Scientific Research (NWO). Additional support came from NWO (SIR-grants SIR 13-3211 and SIR 13-3773), the University of Amsterdam, the ILLC, the European Community (ESPRIT Working Group 8556: Neural and Computational Learning (NeuroCOLT)), the University of Helsinki, Finland and from Shell Nederland B.V.

Cover design by Tobias Baanders.
Printed and bound by Printpartners Ipskamp B.V., Enschede, the Netherlands.

ISBN: 90–5776–009–6

"Nulla pluralitas est ponenda nisi per rationem vel experiantiam vel auctoritatem illius, qui non potest falli nec errare, potest convivi."

(A plurality should only be postulated if there is some good reason, experience, or unfallible authority for it.)

– William of Ockham (c. 1285 – c. 1349)

Acknowledgements

The Beginnings In 1992 I was an undergraduate student of Computer Science specializing in Artificial Intelligence. I was interested in the following question: *can computers learn?* I had attended several classes on 'Machine Learning' but somehow I was not satisfied with what I was told there – until one Thursday afternoon I attended professor Vitányi's lecture on Solomonoff's idea - the idea that learning can be cast in terms of data compression, which, in its practical form, leads to the Minimum Description Length (MDL) Principle. I was sold immediately – this was the most fascinating subject I had ever come across.

The Supervisor In 1994, after obtaining my Master's degree, I became a Ph.D. student at the CWI, privileged to work almost everyday on what was still my favorite subject. Paul Vitányi, the man who had introduced me to the field, became my supervisor. I am very grateful for the trust Paul put in me from the beginning, when he decided to appoint me, a student of computer science, to the mathematics job I was applying for. I want to thank him for four years full of advice in scientific and worldly matters alike. I really enjoyed the freedom he gave me: besides MDL, there was opportunity to work on several other topics as well (yes, I do find other topics interesting too).

The Committee The reading committee consisted of Pieter Adriaans, Krzysztof Apt, Johan van Benthem, Phil Dawid, Michiel van Lambalgen, John-Jules Meyer and Kenji Yamanishi. I want to thank them very much for their critical reading of the manuscript. By writing a thesis of 300 pages I made life difficult not just for myself but also for them!

The Colleagues Of my colleagues at CWI I would particularly like to mention my office mates, former office mates and near–office mates: Jeroen van Maanen, André Berthiaume, Louis Salvail, John Tromp, Dieter van Melkebeek, Ronald de Wolf, Jaap-Henk Hoepman, Lance Fortnow, Leen Torenvliet, Harry Buhrman, Wim van Dam and Barbara Terhal. When suffering from thesis stress, buurman Harry was always there to relieve me by telling me that there was really nothing to worry about. Jeroen had an indirect yet important influence on this thesis – much of the work in Part I addresses questions originally raised by him. About Ronald I shall say more on the next page.

MDL Country In 1997, I spent two months at the University of Helsinki, Finland. The work I did there led to several articles, summarized in Part II of this thesis. I want to thank my hosts and co-authors Petri Kontkanen, Petri Myllymäki, Tomi Silander and chef d'équipe Henri Tirri for what I consider a very successful collaboration which I hope we will continue. Henri Tirri pointed out to me that I really *must* read Jaynes' unfinished book and thus has had a considerable influence on part I of this thesis. The hospitality of Petri and his wife Hanna has made my stay up north very pleasant.

The Crisis In the process of writing a thesis, it is not uncommon to develop one or more crises. In my case, it was a single one, short but severe. I am deeply grateful to Martin Donker, Anne Westhoff and upstairs neighbors Daphne de Leur and Guszti Eiben for supporting me during those dark days! Guszti is also to be thanked for his rôle as a kind of 'older brother' in science, giving me useful advice whenever I needed it.

The Inspirers The scientific environment is not always stimulating: papers get rejected, sometimes with good reason, sometimes without; work is sometimes judged solely on whether it belongs to a certain paradigm and not on its scientific value; people are indifferent to your work while you think it should be of great importance to them, and so on. In such an environment, it is crucial that there are also some people who are simply enthusiastic about what you are doing and who urge you to go on with it – it is indeed partly because of them that for the last four years I actually did go on. Several people helped me in this respect at several times – the first ones that come to mind are scientists Bruno Gaume, Johan van Benthem, Phil Dawid and John-Jules Meijer; and friend, former housemate, co-author and paranymph Mark Steijvers.

The Paranymphs Mark and my other paranymph Mischa Bonn are to be thanked for wanting to perform this job – all the more so since they live in Bloomington, Indiana and in Berlin, respectively.

The Readers Apart from my promotor and the reading committee, there have been two people who really spent quite some time reading preliminary bits of this thesis - the aforementioned Ronald de Wolf and my father, Herbert Grünwald. Ronald eats book chapters for breakfast - it's great to have a roommate whom you can simply tell to read and comment upon a chapter, and he'll just do it – often on the same day. My father, who is not at all acquainted with the subject, nevertheless succeeded in giving lots of useful comments. And this brings me to ...

The Dear Ones - I In a lovely valley in the Austrian alps lies the Sanatorium Grünwald, where tired Ph.D. students have always received a warm welcome. When there, my parents made sure that I could relax completely – they took care of everything. I am immensely grateful for their love and their support.

The Dear Ones - II Finally, of all the people who have helped me with my thesis, there is only one who at times really had to suffer. She understood and she hardly complained; she remained as loving and lovely as ever. I cannot express how grateful I am for that, Louise.

Introduction

To be able to forecast future events, science wants to infer general laws and principles from particular instances. Roughly speaking, this process of *inductive inference* is the central theme in statistics, pattern recognition and the branch of Artificial Intelligence called 'machine learning'. The *Minimum Description Length (MDL) Principle* is a relatively recent method for inductive inference [128]. The fundamental idea behind the MDL Principle is that any regularity in a given set of data can be used to *compress* the data, i.e. to describe it using fewer symbols than needed to describe the data literally. The more regularities there are in the data, the more we can compress it.

Formalization of this idea leads to the concept of *Kolmogorov Complexity* [93], which measures the complexity of a set of data by the length of the shortest computer program that prints the data and then halts. The shortest program for the data is then regarded as the optimal model for the data. By equating 'short' with 'simple', we see that this amounts to a mathematical version of Occam's Razor [93]: if several explanations (programs) of a given phenomenon (data) exist, then we should pick the simplest (shortest) one[1].

The Minimum Description Length Principle

It can be mathematically shown that data compression in terms of Kolmogorov Complexity is almost always the best way to proceed [160]. However, Kolmogorov Complexity is uncomputable and cannot directly be used in practice. The MDL Principle scales down the ideas behind Kolmogorov Complexity in order to make them applicable in practical settings. Rather than considering all computer programs as possible models for the data, one focuses on a class of models $\mathcal{M}$ that is simple enough to allow us to compute, for each model or *hypothesis* H in the class $\mathcal{M}$, how much the data can be compressed on the basis of that model. To describe or equivalently, *encode* the data on the basis of a model H, one first encodes H itself and one then encodes the data with the help of hypothesis H. Choosing the H that minimizes the total code length of the data automatically leads to the selection of a hypothesis that incorporates a trade-off between the complexity and the goodness-of-fit of the hypothesis. The reason for this is that the better a hypothesis fits the data at hand, the more *information* it gives us about the data. The more information we have about the data, the fewer bits we need to encode it. Intuitively, this is because we do not have to

[1] See page v for a formulation that can actually be found in Occam's work ([40], pages 115–117).

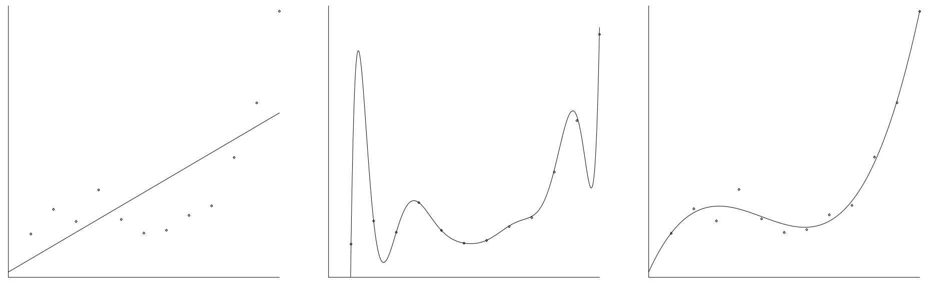

Figure 1: A simple (1.1), complex (1.2) and a trade-off (3rd degree) polynomial.

encode the full data, but only the 'discrepancies' between the hypothesis and the data. It is always possible to find a very complex hypothesis that fits the data extremely well but that does not tell us anything interesting about it. Encoding the data with the help of such a hypothesis takes only very few bits, but to encode (describe) the data we need to describe the hypothesis itself too. The more complex a hypothesis, the more bits we need to describe it. As a result, the gain in description length of the data using a very complex hypothesis is more than offset by the large number of bits needed to describe such a hypothesis. An overly complex hypothesis will therefore not lead to the shortest possible description length. On the other hand, using a very simple hypothesis, the opposite effect occurs: the description length of the hypothesis is very small, but the description length of the data when encoded with the help of the hypothesis will be very high. Assuming that we have 'meaningful' data, that is, it does not consist of purely random noise, the shortest possible description length will usually be attained for relatively simple hypotheses that fit the data reasonably well; according to MDL, these should be preferred both over overly complex hypotheses that have no error at all and over overly simple hypotheses that have very high error. We give a little example. Suppose we want to fit a polynomial to the set of points depicted in Figure 1. Classical linear regression will give us the leftmost polynomial - a straight line that seems overly simple: it does not capture the regularities in the data well. Since for any set of k points there exists a polynomial of the $(k-1)$-st degree that goes exactly through all these points, simply looking for the polynomial with the least error will give us a polynomial like the one in the second picture. This polynomial is overly complex: it reflects the random fluctuations in the data rather than the general pattern underlying it. It seems more reasonable to prefer a simple polynomial with small but nonzero error, as in the rightmost picture. This intuitively right polynomial is exactly the polynomial that the MDL Principle will mechanically give us.

Overview

Here we report the research the author has conducted over the last four years. Most of this research is, either directly or indirectly, related to the MDL Principle. The main research goals were to provide additional justifications for the MDL Principle and to investigate MDL's view on probabilities. We consider both goals in turn.

Justifications of MDL The results of applying the MDL Principle are generally very intuitive. However, mere intuition does not provide enough justification to adopt it for practical use. In this thesis we consider additional justifications for the MDL Principle. Inductive principles like MDL can be justified in different ways. In a theoretical justification, one typically proves that under certain conditions the hypothesis chosen by the principle leads to *provably optimal prediction or classification of future data*. Another theoretical means of justification is to show that a principle can be interpreted as a formal extension of other, existing principles that are already successfully used in practice. A third, practical means of justification is to actually apply the principle in practice and show that this leads to good results. This thesis contains justifications of both theoretical and practical nature.

Probability and Reasoning under Uncertainty Both in the statistical and in the 'reasoning under uncertainty' literature there is fierce debate between those who hold the view that 'all uncertainty should be handled using probability theory' and those who think that probability theory in itself is not enough. In the field of statistics, the two sides in this dispute are occupied by the *Bayesian*[2] or *subjectivist* statisticians on the one hand and the *frequentist* (also called *classical* or *objectivist*) statisticians on the other hand [17]. In the field of reasoning under uncertainty, the discussion is between, once more, the Bayesians on one side, and a menagerie of research communities on the other side; we mention the fields of *fuzzy logic and possibility theory* [41], *certainty factors*, [114], *Dempster-Shafer theory* [138] and *nonmonotonic reasoning* [56].

One of the most controversial aspects of the subjectivist view (actually, it is so controversial that there even are Bayesians who do not accept it[3]) is the *Principle of Insufficient Reason*, advocated as a basis of subjective probability by the great probability theorist P.S. Laplace (1749-1827) [91]:

> If we know that one of k possible alternatives will actually take place, but we have no idea whatsoever whether any of these alternatives is more likely than any other one, then we should assign each of the alternatives probability $1/k$.

The question of whether this makes sense or not has divided philosophers, logicians and statisticians into two camps for over 180 years now [158]. More recently (1957), E.T. Jaynes (1922-1998) extended Laplace's Principle to the 'Principle of Maximum Entropy' [73] which allows one to assign probabilities in the presence of *partial* knowledge. In principle, Maximum Entropy can be used in most if not all of the problems of 'reasoning under uncertainty'. It has actually been applied in lots of practical settings, generally with good results. Nevertheless, it remains as controversial as Laplace's original principle and is still rejected by many [77].

MDL's View on Probabilities It so happens that the notions of 'description method' and 'probability distribution' are very closely connected: every description

[2] Named after Thomas Bayes (1702-1761).

[3] Interestingly, R.T. Cox who provided a very strong justification of the subjectivist viewpoint strongly disagreed with Laplace's principle. See [31], page 31.

method or *code* can be re-interpreted as a probability distribution and vice versa. If data can be coded with the help of model H in only a few bits, then this may be reinterpreted as saying 'the data has a high probability under H - one can define 'probabilities' in this way even for non-probabilistic models like polynomials. From the MDL point of view, a *model* of the data is really a means of describing properties of the data, and hence a 'model' coincides with a 'description method'. Because of the correspondence referred to above, a probability distribution can also be seen simply as a means to describe properties of the data.

This view of probabilities, mainly developed by Rissanen [128], is not so different in principle from the view that is held by some of the Bayesians [36, 38, 77]. What is new is the explicit interpretation of probability distributions in terms of description methods. As discussed informally by Rissanen [128], this interpretation sheds new light on the discussion between the subjectivists and the objectivists. In this thesis, we work out and extend Rissanen's ideas. Briefly, we reach the following conclusion: Probabilities can – and should – indeed be used in many situations where they are not related to frequencies; the 'maximum entropy principle' can indeed be used to assign probabilities in the presence of ignorance. *However, the way such probabilities should be used to arrive at predictions and decisions is different from the way probabilistic knowledge is usually applied.*

Contents of this Work

The thesis consists of three parts. Each part starts with a brief, informal introduction. Part I offers a self-contained introduction to the MDL Principle (chapters 1 and 2) and to Maximum Entropy (Chapter 3). Chapters 4 and 5 introduce some new theoretical concepts concerning MDL. In an Epilogue (page 119) to Part I, it is argued that with the help of these concepts one can identify conditions under which overly simple models can be used unproblematically; this leads to the conclusions about the applicability of probabilistic methods that we sketched above.

In part II, we give an empirical comparison between MDL and some other methods for inductive inference by testing their performance on real-world data. Specifically, in Chapter 6 we compare MDL to the related but distinct Bayesian and maximum likelihood methods of inference. In Chapter 7 we compare MDL to its nearest 'competitor': the Minimum *Message* Length Principle [161, 162].

Part III reports on additional research we have done in the field of 'Reasoning under Uncertainty' within a non-probabilistic paradigm: *non-monotonic temporal reasoning* [56]. Chapter 8 gives an introduction to this field. Chapters 9 and 10 report on the basic research we performed in this field, which is connected to the rest of this thesis through its implicit use of probability, which in turn is implicitly treated from an MDL perspective. In an Epilogue to part III (page 269) we make these connections explicit.

Contents